medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation

Md. Kamrul Hasan[a,b,1,*], Md. Toufick E Elahi[a,b], Md. Ashraful Alam[a,b]

[a] *Department of Electrical and Electronic Engineering (EEE)*
[b] *Khulna University of Engineering & Technology (KUET)*

Abstract

Background and Objective

Although automated Skin Lesion Classification (SLC) is a crucial integral step in computer-aided diagnosis, it remains a challenging task due to inconsistency in texture, color, indistinguishable boundaries, and shapes.

Methods

In this article, we propose an automatic and robust framework for the dermoscopic SLC named Dermoscopic Expert (DermoExpert). The DermoExpert consists of a preprocessing, a hybrid Convolutional Neural Network (hybrid-CNN), and transfer learning. The proposed hybrid-CNN classifier consists of three distinct feature extractors, with the same input images, which are fused to achieve better-depth feature maps of the corresponding lesion. Those distinct and fused feature maps are classified using the different fully connected layers, which are then ensembled to get a final prediction probability. In the preprocessing, we use lesion segmentation, augmentation, and class rebalancing. For boosting the lesion recognition, we have also employed geometric and intensity-based augmentation as well as the class rebalancing by penalizing the loss of the majority class and adding extra images to the minority classes. Additionally, we leverage the knowledge from a pre-trained model, also known as transfer learning, to build a generic classifier, although small datasets are being used. In the end, we design and implement a web application by deploying the weights of our DermoExpert for automatic lesion recognition.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Results

We evaluate our DermoExpert on the ISIC-2016, ISIC-2017, and ISIC-2018 datasets, where our DermoExpert has achieved the area under the receiver operating characteristic curve (AUC) of 0.96, 0.95, and 0.97, respectively. The experimental results outperform the recent state-of-the-art by a margin of 10.0 % and 2.0 % respectively for ISIC-2016 and ISIC-2017 datasets in terms of AUC. The DermoExpert also outperforms, in concerning a balanced accuracy, by a margin of 3.0 % for ISIC-2018 dataset.

Conclusion

Since our framework can provide better-classification on three different test datasets, even with limited training data, it can lead to better-recognition of melanoma to aid dermatologists. Our source code, and segmented masks, for ISIC-2018 dataset, will be made publicly available for the research community for further improvements.

Keywords: Skin lesion classification, Dermoscopic lesion Segmentation, Convolutional neural networks, Transfer learning, Image Augmentation.

1. Introduction

1.1. Problem Presentation

Skin cancer, one in every three cancers worldwide (Ge et al., 2017), is a common type of cancer that originates in the epidermis layer of skin. Ultraviolet radiation exposure is one of the prominent sources of roughly 90 % skin cancer (Narayanamurthy et al., 2018). In 2018, in the United States (US), it is one of the most fifth common diseases in a region under the strong sunshine (Ries et al., 2006). Approximately 2, 490 females and 4, 740 males lost their lives in 2019 (Zhang et al., 2020) due to melanoma, whereas, in the US alone, nearly

*I am corresponding author
Email addresses: m.k.hasan@eee.kuet.ac.bd (Md. Kamrul Hasan), toufick1469@gmail.com (Md. Toufick E Elahi), alam1603001@stud.kuet.ac.bd (Md. Ashraful Alam)
[1]Department of EEE, KUET, Khulna-9203, Bangladesh.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

20 people die from melanoma per day. Age-standardized rates of melanoma of the top 20 countries, as shown in Fig. 1, anticipates the rate of disease that a population would have if it had a standard age structure. The statistical results, in Fig. 1, show that the probability of having melanoma is more in males patients than females. In 2020, roughly 1.0 million new

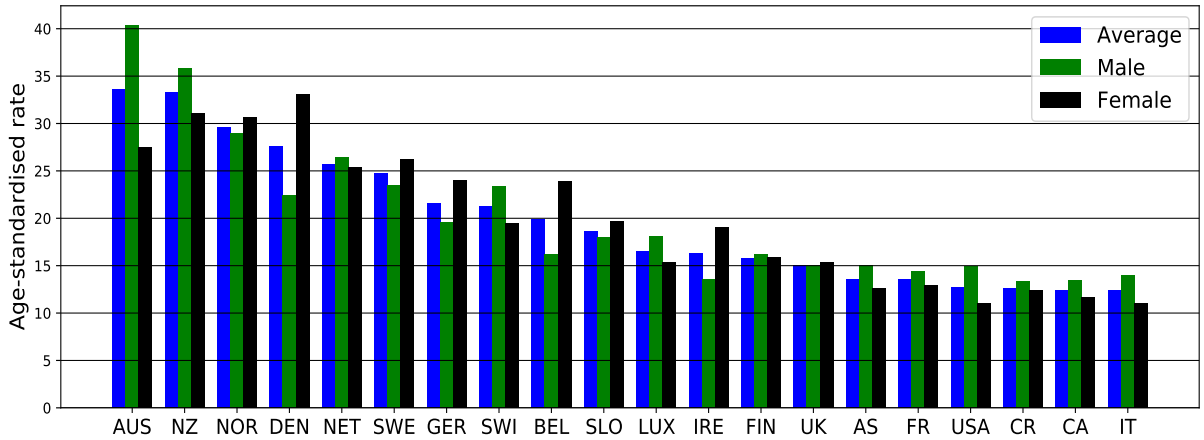


Figure 1: Age-standardized rates per 1.0 million of the top 20 countries with higher melanoma in 2018 (American Institute for Cancer Research, 2018). The bars from left to right are for the decreasing melanoma case of the countries such as Australia (AUS), New Zealand (NZ), Norway (NOR), Denmark (DEN), Netherlands (NET), Sweden (SWE), Germany (GER), Switzerland (SWI), Belgium (BEL), Slovenia (SLO), Luxembourg (LUX), Ireland (IRE), Finland (FIN), United Kingdom (UK) Austria (AS), France (FR), United States of America (USA), Czech Republic (CR), Canada (CA), Italy (IT).

cases of melanoma will recognize. Roughly, $6,850$ new case of deaths due to melanoma are anticipated in 2020, which will contain $4,610$ males and $2,240$ females (Siegel et al., 2020). However, a precise and robust early recognition is very important as the survival rate was as high as apparently $90\,\%$ in advance recognition (Ge et al., 2017). Several imaging techniques, like confocal scanning laser microscopy, optical coherence tomography, ultrasound imaging, and dermoscopic imaging, are currently being used to diagnose skin cancer (Smith and MacNeil, 2011). Dermoscopic images, also known as epiluminescence light microscopy, are most widely used to investigate pigmented skin lesions by the dermatologists (Friedman et al., 1985). Such a visual assessment, via the naked eye, may introduce a faulty-recognition, as it endures from the comparability between the lesions and healthy tissues (Al-Masni et al.,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

2020; Jones et al., 2019). The manual inspection by the dermatologists is often a tedious, time-consuming, and subjective, which can also lead to different recognition results (Al-Masni et al., 2020). However, to mitigate all of the above-motioned limitations and to improve the preciseness in skin cancer recognition, Computer-aided Diagnosis (CAD) has been developed (Jalalian et al., 2017). The classification step, in a CAD system for the SLC, is a crucial component, which is a challenging task due to the presence of diverse artifacts such as markers, body hair & fibers, air bubbles, reflections, on-uniform lighting, rolling lines, shadows, non-uniform vignetting, and patient-specific effects like lesion texture & color, size of affected lesion area (Hasan et al., 2020; Mishraa and Celebi, 2016) as shown in Fig. 2.

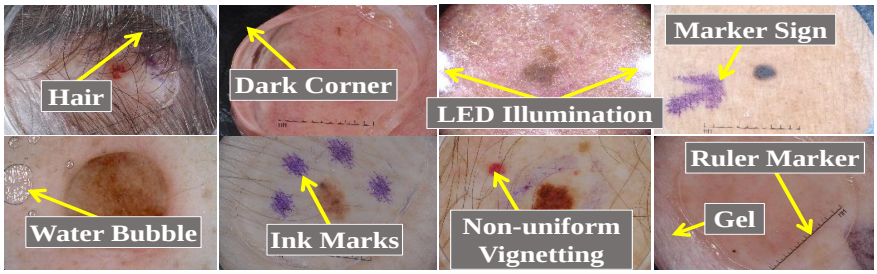


Figure 2: An example of the challenging dermoscopic images in ISIC dataset (Codella et al., 2019, 2018; Gutman et al., 2016; Tschandl et al., 2018) with different artifacts (Hasan et al., 2020).

1.2. Related Works

Nowadays, several methods are being used for the SLC (Brinker et al., 2018; Ma et al., 2015). Yu et al. (2016) had proposed a novel CNN architecture for SLC, where CNN learned from the multiple-image resolutions while leveraging pre-trained CNNs. They constructed a Fully Convolutional Residual Network (FCRN) and enhanced its capability by incorporating a multi-scale contextual information integration scheme. They also integrated their proposed FCRN (for segmentation) and a very deep residual network (for classification) to build a two-stage framework. Majtner et al. (2016a) had presented an automatic melanoma detection system by employing the deep learning method combined with so-called hand-crafted RSurf features (Majtner et al., 2016b) and Local Binary Patterns (Ahonen et al.,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

2006). Finally, they had used Support Vector Machine (SVM) (Furey et al., 2000) for the classification of the skin lesion. Transfer learning in VGGNet architecture (Simonyan and Zisserman, 2014) was implemented by Lopez et al. (2017), where they compared the classification results with the CNN training from scratch. A novel multi-tasking deep learning model based on GoogleNet (Szegedy et al., 2015) and UNet (Ronneberger et al., 2015) was proposed by Yang et al. (2017) for concurrent segmentation and classification, where they passed both the RGB and HSV dermoscopic images as an input. Quang et al. (2017) had presented a fully encoder-decoder architecture for segmentation, which was then classified using the VGG16. Structural Co-occurrence Matrix (Zucker and Terzopoulos, 1980) of the main frequencies were extracted from dermoscopic images and classified using the multilayer perceptron by Rebouças Filho et al. (2018), where they employed exhaustive preprocessing steps. A framework for dermoscopic image recognition via both a deep learning method and a local descriptor encoding strategy was proposed by Yu et al. (2018). The deep representations of a rescaled dermoscopic images were extracted via a very deep Residual neural Network (ResNet) (He et al., 2016a), which were then aggregated by the orderless visual statistic features based on Fisher Vector (FV). Finally, the FV encoded representations were used to classify melanoma images using an SVM with a Chi-squared kernel (Li et al., 2013). Rehman et al. (2018) performed segmentation using the intensity thresholding on each channel. They used CNN for feature extraction, which was then classified using the neural network-based classifier. A classification of skin lesions using a deep CNN (Inception-V4) (Szegedy et al., 2017), data augmentation, and traditional machine learning classifiers were proposed by Pham et al. (2018). They also performed both geometric and color augmentation, and then deep CNN was used as a feature extractor. Finally, those features were classified using SVM, Random Forest (Liaw et al., 2002), and Neural Network (Beale et al., 1996). A multi-channel-ResNet was proposed by Guo and Yang (2018), where the original images were passed through several channels. Each channel of which corresponds to a preprocessing method, which was concatenated to obtain the final probability. Transfer learning strategy on VGG16 and GoogLeNet architectures were employed by Majtner et al. (2018), where a preprocessing, based on image augmentation such as horizontal flipping &

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

rotation and color normalization, was also used to build a generic network. A deep learning framework, consisting of two FCRNs, was proposed by Li and Shen (2018) to get segmentation and coarse classification results concurrently. They also developed a Lesion Index Calculation Unit (LICU) to calculate the distance heat-map, where the coarse classification result was refined according to that generated distance map. Brinker et al. (2019) used transfer learning in ResNet-50 architecture to classify the skin lesion in a very generic manner. For doing so, they performed three techniques such as exclusive training of the adapted last layer, fine-tuning the parameters of all layers, and sudden increases of the learning rate at specific time steps during fine-tuning. Mahbod et al. (2019) proposed an ensemble scheme by combining intra- and inter-architecture network fusion, where they used fine-tuning of AlexNet (Krizhevsky et al., 2012), VGGNet, and two variations of ResNets. The final prediction was accomplished using the SVM, where the average classification vectors from different sets were fused. Akar et al. (2019) had designed and implemented a cloud-based SLC system that consists of a CNN-based classifier and a mobile application that runs on Android and iOS. Zhang et al. (2019) had proposed an Attention Residual Learning CNN (ARL-CNN) model for the SLC, which was composed of multiple ARL blocks, a global average pooling layer, and a classification layer. Each ARL block jointly used residual and novel attention learning mechanisms to improve its ability for discriminating representation. Instead of using extra learnable layers, they proposed an attention learning mechanism to exploit the intrinsic self-attention ability of DCNNs. Amin et al. (2020) had performed the segmentation using the 2-D wavelet transform and Ostu algorithm, where they converted the RGB images to Lab images. They also extracted deep features using AlexNet and VGG16 model, which were selected using the Principle Component Analysis (PCA). Finally, they obtained the classification results by employing the k-nearest Neighbour (k-NN) (Cunningham and Delany, 2007) and SVM. Al-Masni et al. (2020) had employed the segmentation and classification on the segmented images, where they employed four different CNN networks such as Inception-V3 (Szegedy et al., 2016), ResNet-50, Inception-ResNet-V2 (Szegedy et al., 2017), and DenseNet-201 (Huang et al., 2017) for the classification. They performed different geometric augmentation and transfer learning. The stacking ensemble method based

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

on the meta-learning algorithm was proposed by Ghalejoogh et al. (2020), where two new hybrid approaches were also introduced to combine the heterogeneous classifiers. The effect of dermoscopic image size based on pre-trained CNNs and transfer learning was analyzed by Mahbod et al. (2020), where they resized from 224×224 to 450×450. Three well-established CNNs such as EfficientNetB0, EfficientNetB1, and SeReNeXt-50 were explored for the SLC. They had also proposed and evaluated a multi-scale multi-CNN (MSM-CNN) fusion approach based on a three-level ensemble strategy that utilizes the three network architectures trained on cropped dermoscopic images of different scales. An architecture search framework was presented by Kwasigroch et al. (2020) to detect the malignant melanoma. The hill-climbing search strategy was employed along with network morphism operations to explore the search space for finding a suitable network structure. Such morphism operations allow the increment of network size with the use of the previously trained network. Valle et al. (2020) optimized the hyperparameter of the deep CNN models like ResNet-101-V2 (He et al., 2016b) and Inception-V4 employing transfer learning and segmentation. They enhanced the invariance properties of the network utilizing the image augmentation. They also performed extensive experiments to select the best performing classifier using the ANOVA test. Finally, the authors concluded that the transfer learning and ensembling model is a better choice for designing the SLC systems. A Multi-class Multi-level (MCML) algorithm was proposed by Hameed et al. (2020), which was based on "divide and conquer" method for the SLC. The MCML consists of four integral parts: pre-processing, segmentation, feature extraction, and classification. They used traditional machine learning and deep learning models as classifiers. However, their pipeline highly dependent on preprocessing like hair, black frame, and circle removing. Gessert et al. (2020) ensembled several deep learning methods, including EfficientNets, SENet (Hu et al., 2018), and ResNeXt (Mahajan et al., 2018), utilizing a selection strategy. They used multi-resolution input by cropping the original images. The authors also fed the metadata as a feature vector, which was then concatenated with the CNN models. Khan et al. (2020) developed a framework for the SLC, which consists of the localization of lesion ROI via faster region-based CNN, feature extraction, and feature selection by iteration-controlled Newton-Raphson method. Firstly,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

an artificial bee colony method was used for the contrast stretching, which was then used for the lesion segmentation. The DenseNet-201 was then utilized to extract the deep features. Finally, the authors employed a multilayer perceptron as a classifier. A Generative Adversarial Network (GAN) was employed by Qin et al. (2020) for the generation of high-quality synthetic skin images. A ResNet-50, with pre-trained ImageNet weights, was used as a lesion classifier. To classify skin lesion, Chaturvedi et al. (2020) developed a diagnosis system based on MobileNet (Howard et al., 2017) using transfer learning. They used different geometric augmentations such as rotation, zoom, horizontal and vertical flip. Almaraz-Damian et al. (2020) proposed a pipeline, which consists of different integral steps such as preprocessing, feature extraction, feature fusion, and classification. As a preprocessing, authors extracted ROIs of the lesion, where they also enhanced the intensity of the extracted ROIs. They extracted different handcraft features like shape, color, texture, and CNN features, where mutual information was employed to extract CNN features. To classify the ROIs, several classification methods such as Linear Regression (LR), SVM, and Relevant Vector Machines (RVMs) were used. As a preprocessing, Mporas et al. (2020) used a median filter, which was followed by bottom-hat filtering to detect the hair or similar noise. The segmentation was performed to extract lesion ROIs on the grayscale image using the active contour model. Finally, different color based features were extracted, which were then classified using multilayer perceptron and different machine learning algorithms. Yilmaz and Trocan (2020) compared the performance of deep CNN such as AlexNet, GoogLeNet, and ResNet-50 for the classification of skin lesions. The authors experimentally demonstrated that ResNet-50 was the best performing classifier, whereas AlexNet was better for time complexity. Pereira et al. (2020) used the gradient-based histogram thresholding and local binary pattern clustering for the border-line characteristics of skin lesions. Then, border-line characteristics are concatenated with CNN to boost the lesion classification performance.

1.3. Our Contributions

In this literature, we propose a robust and generic framework for the dermoscopic SLC, called DermoExpert, where the preprocessing, the proposed hybrid-CNN classifier, and the

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

transfer learning are the integral steps. The proposed hybrid-CNN classifier uses a two-level ensembling. In the first level, we have concatenated, in channel-wise, 2D feature maps of different feature map generators to enhance the depth information. The concatenated or fused feature map and individual feature maps are trained together. Finally, we have aggregated the different outputs at the second level. In the proposed preprocessing, we employ a lesion segmentation, augmentation, and rebalancing, where the ROIs segmentation enables the classifier to learn only the features from the specific lesion areas while avoiding the surrounding healthy tissues. The precise segmentation, with less-coarseness, is a critical prerequisite step for the classification as it extracts abstract region and detailed structural description of various types of skin lesions. The segmentation has performed by using our recent state-of-the-art DSNet (Hasan et al., 2020) after fine-tuning with other ISIC datasets. We use geometry and intensity-based image augmentation, as a preprocessing, to increase the number of training images and overcome the overfitting. We also use a class rebalancing by adding extra images from the ISIC archive and weighting the loss function to tackle the unwanted biasing toward the majority class. Moreover, the transfer learning, by leveraging the weights from a pre-trained model, was applied to all the distinct feature extractors in the proposed classifier. To our best knowledge, our proposed DermoExpert has achieved state-of-the-art results on the three IEEE International Symposium on Biomedical Imaging (ISBI) datasets such as ISIC-2016, ISIC-2017, and ISIC-2018 with respectively two, three, and seven types of skin lesion, while being an end-to-end system for the SLC. Additionally, we have implemented and compared the performance of our DermoExpert against several well-established deep learning classifications approaches like Xception, ResNet, and DenseNet under the same experimental environments and preprocessing using the same dataset. At last, we have implemented a web application by deploying the trained DermoExpert for the clinical application, which runs in a web browser.

The remaining sections of the article are organized as follows: section 2 explains the proposed methodologies and datasets, where we explicitly mention the extensive experiments. Section 3 describes the obtained results from different extensive experiments. The results are explained with a proper interpretation in section 4, where we also present a user appli-

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

cation by employing the trained model for the SLC. Finally, section 5 concludes the article with future works.

2. Methods and Materials

This section focuses on methods and materials used for the study, where subsection 2.1 explains the proposed DermoExpert and datasets for the SLC. In subsection 2.2, we present the metrics for evaluation and hardware used to conduct this research. Subsection 2.3 presents the training protocol and experimental details.

2.1. Proposed Framework

The proposed DermoExpert, as shown in Fig. 3, consists of dermoscopic image preprocessing, transfer learning, and proposed hybrid-CNN classifier, where we have employed lesion segmentation for ROI extraction, augmentation, and rebalancing the target classes as a preprocessing. We have validated our DermoExpert with three different datasets having a different number of target classes to provide the evidence of generality and versatility of the proposed DermoExpert. In the proposed DermoExpert, there are three distinct inputs for three different datasets, which are termed as I_1, I_2, and I_3. The input I_1, I_2, or I_3 generates the corresponding output O_1, O_2, or O_3 using the proposed preprocessing and hybrid-CNN classifier. The different integral and crucial parts of the DermoExpert are briefly described as follows:

2.1.1. Datasets

The three well-known datasets of skin lesions such as ISIC-2016 (Gutman et al., 2016), ISIC-2017 (Codella et al., 2018), and ISIC-2018 (Codella et al., 2019; Tschandl et al., 2018) are used, in this research, to evaluate the proposed DermoExpert (see Fig. 3), which are respectively denoted by I_1, I_2, and I_3 as an input to the DermoExpert. Table 1 shows the class-wise distributions and short descriptions of the used datasets. The ISIC-2016 is a binary classification task to classify the lesions as either Nevus (Nev) or Melanoma (Mel). The ISIC-2017 and ISIC-2018 are the multi-class classification tasks. In the ISIC-2017, the

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

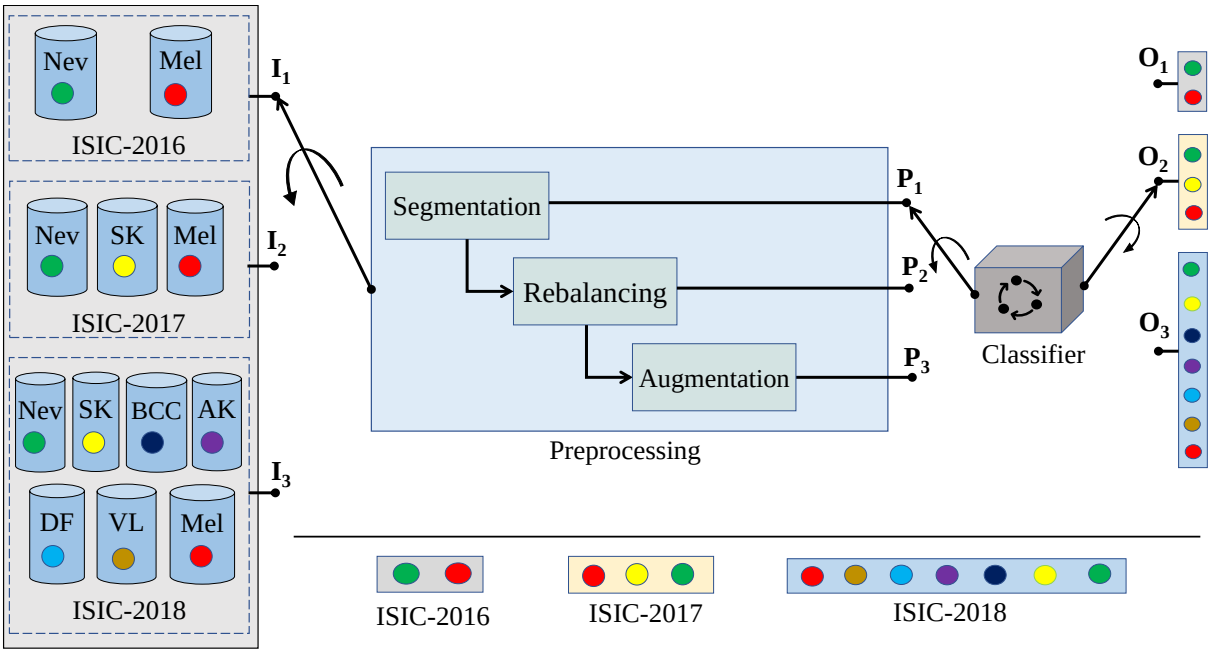


Figure 3: The proposed DermoExpert for the SLC, where the preprocessing has incorporated with the pipeline to build a precise and robust diagnostic system. The input I_1, I_2, or I_3 is followed by the preprocessing (P_1, P_2, or P_3) and then by the proposed classifier to generate the corresponding output O_1, O_2, or O_3.

lesion requires to classify as either Nevus (Nev) or Seborrheic keratosis (SK) or Melanoma (Mel). The ISIC-2018 comprises of Nevus (Nev), Seborrheic keratosis (SK), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Dermatofibroma (DF), Vascular Lesion (VL), and Melanoma (Mel) classes. However, the ground-truths of validation and test images are not provided for the ISIC-2018 dataset. We have applied a cross-validation technique for ISIC-2018 dataset to select training, validation, and testing images. The resolutions, in pixels, of all 8-bit dermoscopic images, as in Table 1, are 540×576 to 2848×4288, 540×576 to 4499×6748, and 450×600 respectively for ISIC-2016, ISIC-2017, and ISIC-2018 datasets. The class-distribution of all ISIC datasets, as shown in Table 1, demonstrates that the images are imbalanced, which makes the classifier to be biased to the particular class having more samples. However, we have employed several techniques for rebalancing to build a generic classifier for the skin lesion diagnosis even though datasets are imbalanced.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Table 1: A concise description and class-distribution of the ISIC-2016, 2017, and 2018 datasets.

SL #	Class Types	Description	Train	Validation	Test
01	Nevus (Nev)	Benign tumor derived from melanocytic	727	*	304
02	Melanoma (Mel)	Malignant tumor derived from melanocytic	173	*	75
Total Images in ISIC-2016			900	*	379
01	Nevus (Nev)	Benign tumor derived from melanocytic	1372	78	393
02	Seborrheic Keratosis (SK)	Benign tumor derived from non-melanocytic	254	42	90
03	Melanoma (Mel)	Malignant tumor derived from melanocytic	374	30	117
Total Images in ISIC-2017			2000	150	600
01	Nevus (Nev)	Benign tumor derived from melanocytic	6705	*	*
02	Seborrheic Keratosis (SK)	Benign tumor derived from non-melanocytic	1099	*	*
03	Basal Cell Carcinoma (BCC)	Benign tumor derived from nonmelanocytic	514	*	*
04	Actinic Keratosis (AK)	Benign tumor derived from keratinocytes	327	*	*
05	Dermatofibroma (DF)	Benign tumor derived from histiocytes	115	*	*
06	Vascular Lesion (VL)	Benign tumor derived from blood vessel cell	142	*	*
07	Melanoma (Mel)	Malignant tumor derived from melanocytic	1113	*	*
Total Images in ISIC-2018			10015	*	*

*: Not available publicly

2.1.2. Proposed Preprocessing

The proposed preprocessing, as shown in Fig. 3, consists of segmentation, augmentation, and class rebalancing, which are briefly described as follows:

Segmentation. The segmentation, to separate homogeneous lesion areas, is the critical component for diagnosis and treatment pipeline (Hesamian et al., 2019). It is also a crucial prerequisite for the skin lesion diagnosis as it extracts promising skin lesion features and delivers critical information about the shapes and structures. A recent state-of-the-art DSNet (Hasan et al., 2020), for dermoscopic skin lesion (ISIC-2017) segmentation, has been used as a lesion ROI extractor. We fine-tune the adopted DSNet with the ISIC-2016 and ISIC-2018 datasets to extract the lesion ROI of all three datasets as it was trained and tested on only the ISIC-2017 dataset (Hasan et al., 2020).

Augmentation. CNN's are heavily reliant on big data to avoid overfitting. Unfortunately, many application domains, such as automation in lesion disease diagnosis, suffer from a small size as a huge number of manually annotated training images are not yet available (Harangi, 2018). Augmentation is a very potential preprocessing for training the deep

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

learning models as they are highly discriminative (Hussain et al., 2017). Data augmentation encompasses a technique that enhances the size and quality of training datasets to build a better-CNN classifier (Shorten and Khoshgoftaar, 2019). The geometric-based augmentation such as rotation (around $row/2$ and $col/2$) of 180° and 270°, and the intensity-based augmentation such as gamma, logarithmic, & sigmoid corrections, and stretching, or shrinking the intensity levels are employed in the proposed preprocessing. The gamma correction with two gammas (γ) of 0.7 and 1.7 has performed to change the luminance of the dermoscopic images by $I_{out} = I_{in}^{\gamma}$, where I_{out} and I_{in} are the output and input luminance values. The logarithmic correction, for enhancing an image that provides better contrast and a more structural detailed image, has employed by $I_{out} = G \times \log(1 + I_{in})$, where $G = 0.5$ and I_{out} & I_{in} are the gain and input & output images respectively. We have employed sigmoid correction by $I_{out} = \frac{1}{1+\exp^{G\times(C-I_{in})}}$, where $G = 15$, $C = 0.4$, and I_{out} & I_{in} are the multiplier in exponential's power, cutoff of the function that shifts the characteristic curve in horizontal direction, and input & output images respectively. We also stretch or shrink the intensity levels between the minimum and the maximum intensities.

Rebalancing. All of the three dermoscopic datasets are imbalanced, as presented in Table 1, which means that the number of images in the classes of each dataset is not the same. This scenario is quite common in the medical diagnosis field as manually annotated training images are not sufficiently available (Harangi, 2018), where the positive cases are the minority as compared to negative cases. The unwanted biasing toward the majority class likely to happen in the case of supervised CNN classifiers. However, in this research, we have added additional images from the ISIC archive website (ISIC, 2018), and we also penalize the majority class by weighting the loss function. Such a weighting pays more attention to samples from the minority class. Here, we increase the weight of samples from underrepresented classes with a factor of $W_j = N/N_j$, where W_j, N, and N_j are the weight for class j, the total number of samples, and the number of samples in class j respectively.

Preprocessing Employment. We have sub-divided the preprocessing (P) into the segmentation (P_1), segmentation & class rebalancing (P_2), and segmentation, class rebalancing, & augmentation (P_3) to investigate their effects in the proposed DermoExpert for

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

the SLC.

2.1.3. Proposed CNN-based Classifier and Transfer Learning

Deep CNNs are an excellent feature extractor that can avoid complicated and expensive feature engineering. It has achieved tremendous success since 2012 (Rawat and Wang, 2017). Sometimes it rivals human expertise. For example, in CheXNet (Rajpurkar et al., 2017), a CNN, having 121 layers, was trained on more than 1.0 million chest X-rays. It was capable of achieving a better performance than the four experts. Additionally, Kermany et al. (2018) used Inception-V3 (Szegedy et al., 2016), where the training set contained roughly 1.0 million optical coherence tomographic images. They compared the results with six radiologists, where experts got high sensitivity but low specificity, while the CNN-based system got high values of balanced accuracy. However, those above methods are blessed with a huge number of labeled images, which are hard to collect as it needs a lot of professional expertise for annotation (Yadav and Jadhav, 2019). Moreover, individual CNN architecture may have different capabilities to characterize or represent the image data, which is often linked to a network's depth (Kumar et al., 2016). However, CNN's maybe indirectly limited when employed with highly variable and distinctive image datasets with limited samples such as dermoscopic image datasets (Codella et al., 2019, 2018; Gutman et al., 2016; Tschandl et al., 2018). In this context, we propose a hybrid-CNN classifier by leveraging several core techniques of the current CNN networks to build a generic skin lesion diagnostic system with limited images for training. Fig. 4 depicts the proposed hybrid-CNN classifier. In our hybrid-CNN, the input batch of images is simultaneously passed through three different Feature Map Generators (FMGs) to obtain different presentations of the feature maps (see the output of the encoders in Fig. 4). The proposed hybrid-CNN classifier comprises of the following integrated parts:

Feature Map Generators-1 (FMG-1). The FMG-1 (f^1) consists of the identity or residual and convolutional blocks (He et al., 2016a), where the skip connections allow the information to flow or skip. Fig. 5 shows the constructional details of the identity and convolutional blocks. The skip connection in the residual blocks, as in the proposed hybrid-

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

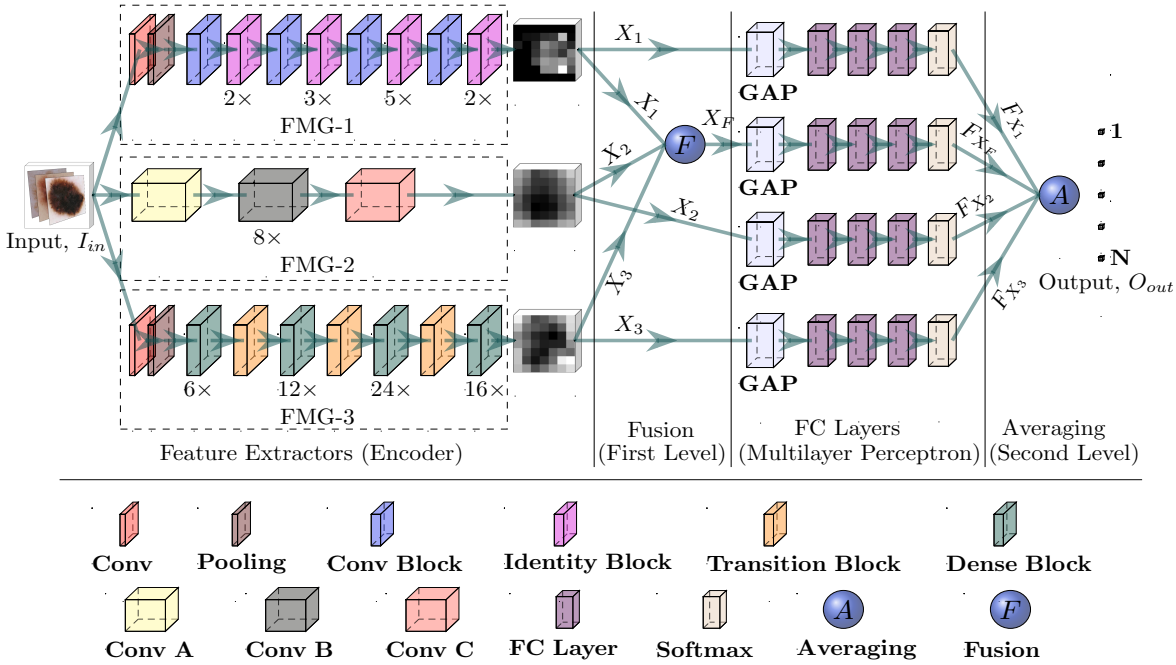


Figure 4: The proposed hybrid-CNN classifier, where three distinct feature extractors, also called an encoder, receives the same input image (I_{in}). The first step, encoders, are followed by the second step, called a fusion (first level ensembling). Then, the third step, called the FC layer, is followed by the fourth step, called an averaging (second level ensembling), to get the final output (O_{out}).

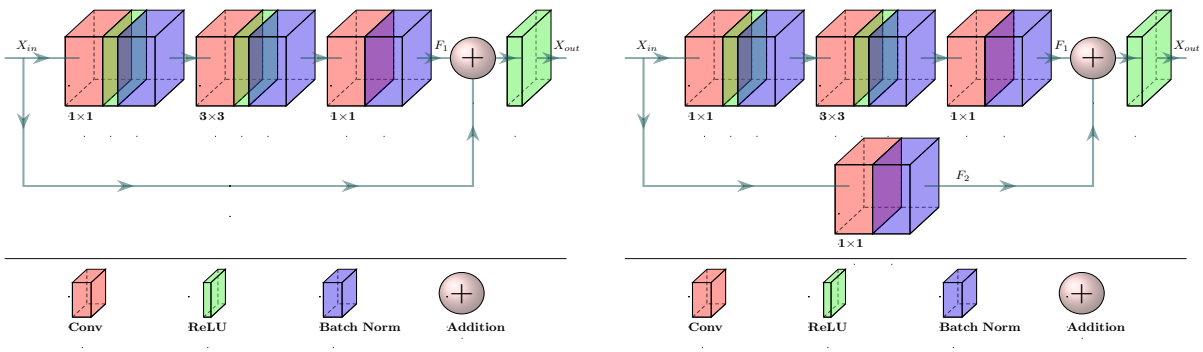


Figure 5: The residual (or identity) (left) and convolutional (right) blocks (He et al., 2016a) of the FMG-1. The output (X_{out}) is the summation of X_{in} and the process (F), where $X_{out} = F_1(X_{in}) + X_{in}$ for residual block and $X_{out} = F_1(X_{in}) + F_2(X_{in})$ for convolutional block.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

CNN classifier, will have two benefits: firstly, the new layers will not hamper the performance as regularisation will skip over them, and secondly, if the new layers are useful, even with the presence of regularisation, the weights or kernels of the layers will be non-zero. However, in FMG-1, a 7×7 input convolution, followed by a max-pooling with a stride of 2, and a pool size of 3×3, is used before the identity and convolutional blocks. By stacking these blocks on top of each other (see Fig. 4), an FMG-1 has formed to get the feature map, where the notation ($n\times$) under the identity block denotes the number of repetitions (n times). The output feature map from the FMG-1 is defined as $X_1 = f^1(I_{in})$, where $X_1 \in \mathcal{R}^{B\times H\times W\times D}$, and B, H, W, and D respectively denote the batch size, height, width, and depth (channel).

Feature Map Generators-2 (FMG-2). The FMG-2 (f^2) consists of the entry flow (Conv A), middle flow (Conv B), and exit flow (Conv C) blocks, which were originally proposed by Chollet (2017). The constructional details of those blocks are given in Fig. 6. The batch of images first go through the entry flow, then through the middle flow, which is repeated eight times ($8\times$), and finally through the exit flow. All the flows, as in the proposed hybrid-CNN classifier, have used depth-wise separable convolution and residual connection. The former one is used to build a lightweight network, whereas the latter one for the benefits, which were mentioned earlier. The output feature map from the FMG-2 is defined as $X_2 = f^2(I_{in})$, where $X_2 \in \mathcal{R}^{B\times H\times W\times D}$, and B, H, W, and D respectively denote the batch size, height, width, and depth (channel).

Feature Map Generators-3 (FMG-3). The remaining FMG-3 (f^3) consists of the dense and transition blocks (Huang et al., 2017), where the constructions of those blocks are shown in Fig. 7. The FMG-3, as in the proposed hybrid-CNN classifier, gets rid of the requirement of learning repetitive features, which can learn the absolute features of the skin lesion (Hasan et al., 2020). The feature re-usability of FMG-3 reduces the vanishing-gradient problem as well as it strengthens the feature propagation (Huang et al., 2017). It also enables the convolutional layer to access the gradients of all the previous layers by using a skip connection as depicted in Fig. 7. However, the output feature map from the FMG-3 is defined as $X_3 = f^3(I_{in})$, where $X_3 \in \mathcal{R}^{B\times H\times W\times D}$, and B, H, W, and D respectively denote the batch size, height, width, and depth (channel).

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

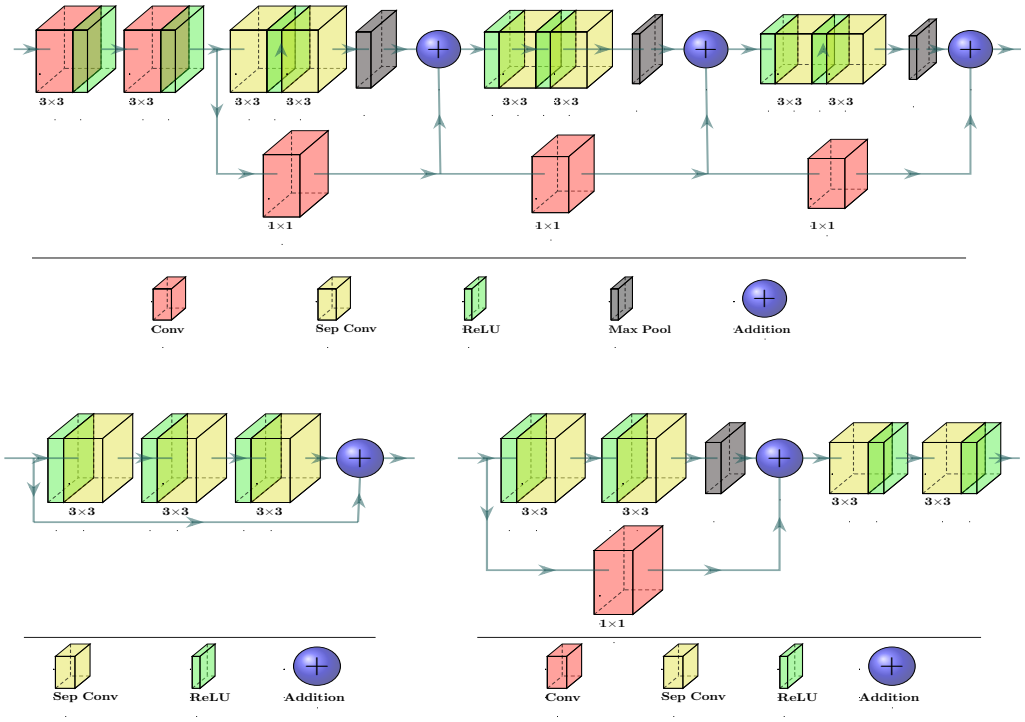


Figure 6: The entry flow (Conv A) (top), middle flow (Conv B) (bottom-left), and exit flow (Conv C) (bottom-right) blocks (Chollet, 2017) of the FMG-2, where depth-wise separable convolutions were employed in lieu of traditional convolutions to make it lightweight for real-time applications.

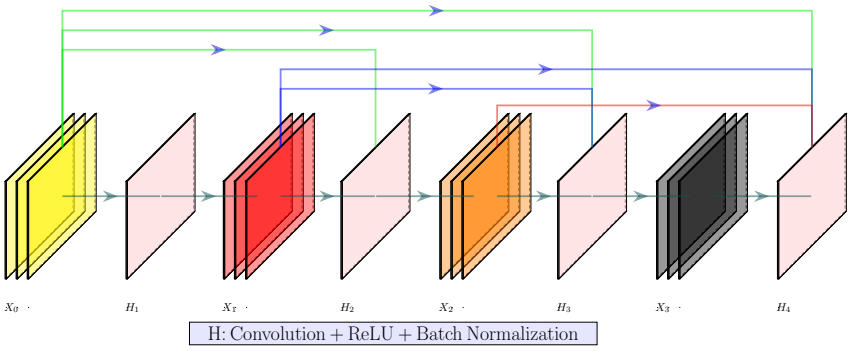


Figure 7: Typical dense block of the FMG-3 for a growth rate of 3.0. Each nth layer of a dense block accepts the feature maps of all previous convolutional layers. The mathematical expression (Huang et al., 2017) of such a reusibiility is $X_n = H_n([X_0, X_1, X_2, ..., X_{n-1}])$, where H_n is the composite function (Huang et al., 2017) of the nth layer, which consists of a convolution, ReLU, and batch normalization.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Fusion (First level ensembling). We have employed those three FMGs, to get the distinct feature maps (see in Fig. 8 (c), Fig. 8 (d), and Fig. 8 (e)) for building a proposed hybrid-CNN classifier. In doing so, it is very impractical, in the beginning, to select the

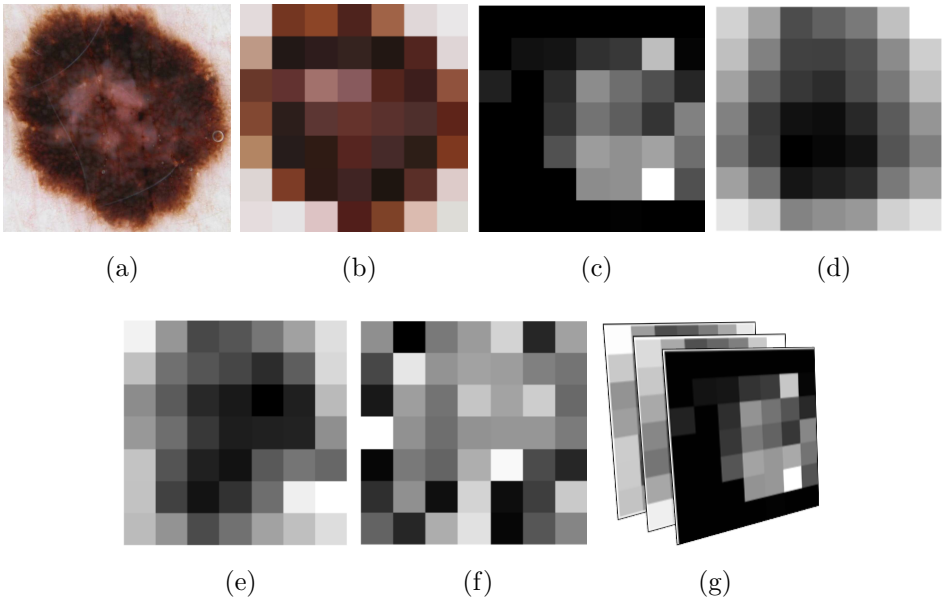


Figure 8: The extracted feature map from different FMGs with corresponding original input image (a), where (b) the down-sampled (7×7) image using nearest-neighbor interpolation, (c) the output of FMG-1, (d) the output of FMG-2, (e) the output of FMG-3, (f) the average of the three FMGs, and (g) the channel-wise concatenation of the output of the three FMGs.

fusion mechanism as there are many ways to get the better-fused map (X_F). However, we perform two types of fusion such as fusion by channel-concatenation (see Fig. 8 (g)) and fusion by channel-averaging (see Fig. 8 (f)), and we have named them as a first-level ensembling. In channel-concatenation, the fused feature map is $X_{FC} = [X_1 +\!\!+ X_2 +\!\!+ X_3]$, where $X_{FC} \in \mathcal{R}^{B \times H \times W \times 3D}$ and $+\!\!+$ denotes the channel concatenation. In channel-averaging, the fused feature map is $X_{FA} = \frac{1}{N=3} \sum_{n=1}^{N=3} X_n$, where $X_{FA} \in \mathcal{R}^{B \times H \times W \times D}$, $\sum$ is the element-wise summation, and N is the numbers of FMG.

Fully Connected Layer. The different feature maps are classified into desired categories using the Fully Connected (FC) layers, where the output is denoted by F_M for M^{th} input feature map. However, to vectorize the 2D feature maps into a single long continuous

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

linear vector before the FC layer, we use a Global Average Pooling (GAP) layer (Lin et al., 2013), which improves generalization and prevents overfitting. The GAP layer performs a more extreme dimensionality reduction to avoid overfitting. An $height \times weight \times depth$ dimensional tensor, in GAP, is reduced to a $1 \times 1 \times depth$ vector by transferring $height \times width$ feature map to a single number. Such a GAP layer contributes to the lightweight design of the CNN classifiers. Additionally, each FC layer is followed by a Dropout layer (Srivastava et al., 2014) as a regulariser, where we randomly set 50.0 % neurons of the FC layer to zero during the training. Such a Dropout layer can build a generic CNN classifier by reducing the overfitting.

Averaging (Second level ensembling). Finally, the output probability ($O_{j=1,2,3}$) is the average of the outputs of different F_M, and we have named it as a second level ensembling. The output ($O_{j=1,2,3}$) lies in N-dimensional space, where $O_1 \in \mathcal{R}^{N=2}$, $O_2 \in \mathcal{R}^{N=3}$, and $O_3 \in \mathcal{R}^{N=7}$ respectively for the inputs I_1, I_2, and I_3 (see subsection 2.1 and Fig. 3) by applying the proposed preprocessing (either P_1 or P_2 or P_3).

Possible hybrid-CNN classifiers. However, in this literature, we propose five possible ensembling classifiers by using first level and second level ensembling, which are enlisted as follows:

1. **Method-1:** Selection of only the first-level ensembling by using the fused feature map X_{FC}, and performing the classification using an FC layer ($F_{X_{FC}}$) for achieving a final probability O_j

2. **Method-2:** Selection of only the first-level ensembling by using the fused feature map X_{FA}, and performing the classification using an FC layer ($F_{X_{FA}}$) for achieving a final probability O_j

3. **Method-3:** Selection of only the second-level ensembling by employing the feature maps (X_1, X_2, and X_3) except the fused maps (X_{FC} and X_{FA}), and performing the classification using the FC layers (F_{X_1}, F_{X_2}, and F_{X_3}) for achieving a final probability O_j

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

4. **Method-4:** Selection of both the first and second-level ensembling by using the feature maps (X_1, X_2, X_3, and X_{FA}) except the fused map (X_{FC}), and performing the classification using the FC layers (F_{X_1}, F_{X_2}, F_{X_3}, and $F_{X_{FA}}$) for achieving a final probability O_j

5. **Method-5:** Selection of both the first and second-level ensembling by using the feature maps (X_1, X_2, X_3, and X_{FC}) except the fused map (X_{FA}), and performing the classification using the FC layers (F_{X_1}, F_{X_2}, F_{X_3}, and $F_{X_{FC}}$) for achieving a final probability O_j

However, we perform the ablation studies (see the subsection 3.3) on the above-mentioned five different methods to obtain the best hybrid-CNN classifier for the SLC.

Transfer Learning. Moreover, when the number of data is relatively small, as the skin lesion datasets in this literature, the model starts to overfit after several epochs. However, the scarcity of such a relatively small medical image datasets has been partially overcome by employing a transfer learning (Shin et al., 2016; Tajbakhsh et al., 2016). It uses the representations learned by a previous model and employs a new domain, which also reduces the need for large computational power (Talo et al., 2019). However, in the proposed DermoExpert, we use the previously trained weights to our FMG-1, FMG-2, and FMG-3 for transferring the knowledge.

2.2. Hardware and Evaluation

Hardware. We have implemented our DermoExpert on a *Windows*-10 machine using the Python programming language with different Python and Keras (Géron, 2019) APIs. The hardware configuration of the used machine are: Intel® Core™ i7-7700 HQ CPU @ 2.80 GHz processor with Install memory (RAM): 16.0 GB and GeForce GTX 1060 GPU with 6 GB GDDR5 memory.

Evaluation Metrics. We have evaluated the results using several metrics, where each metric evaluates the model from different aspects. We use recall, specificity, and intersection over union (IoU) for measuring the segmentation performance quantitatively as an accurate

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

mask is the crucial prerequisite of the proposed DermoExpert. The recall and specificity respectively measure the percentage of true and wrong regions, whereas the IoU measures the amount of the overlapping between the true and predicted masks. We evaluate the proposed DermoExpert using the recall, precision, and F1-score. The recall quantifies the type-II error (the lesion, with the positive syndromes, inappropriately fails to be nullified), and precision quantifies the positive predictive values (percentage of truly positive recognition among all the positive recognition). The F1-score indicates the harmonic mean of recall and precision, which shows the tradeoff between them. We have also reported the confusion matrix for evaluating the DermoExpert by investigating the class-wise performance of the SLC. Moreover, the Receiver Operating Characteristics (ROC) with Area Under the ROC Curve (AUC) value is also used to quantify the prediction probability of any randomly picked sample.

2.3. Training Protocol and Experiments

As we segment the lesion using the recent state-of-the-art DSNet, the kernels, in both the encoder and decoder, have been initialized with the pre-trained weights of the DSNet. We resize all the images to 192×256 pixels using the nearest-neighbor interpolation, for the segmentation, as the DSNet receives the images having a resolution of 192×256 pixels. Additionally, we have standardized and rescaled the training images to $[0\,1]$. The fine-tuning of the DSNet has been performed using the following loss function (L) (Hasan et al., 2020) as Eq. 1.

$$L\left(y, \hat{y}\right) = 1 - \frac{\sum_{i=1}^{N} y_i \times \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i - \sum_{i=1}^{N} y_i \times \hat{y}_i} - \frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \quad (1)$$

where y and $\hat{y}$, N respectively denote the true and predicted label, the total number of pixels.

The pre-trained weights from ImageNet (Deng et al., 2009) were applied to initialize the kernels of all the FMGs. Xavier distribution, also called glorot uniform distribution (Glorot

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

and Bengio, 2010), is used to initialize the kernels in FC layers. It draws the samples from a truncated normal distribution centered on 0.0 with a standard deviation of $\sqrt{(}2/(F_{in}+F_{out}))$. F_{in} and F_{out} respectively denotes the number of input and output units in the weight tensor. The aspect ratio distribution extracted ROIs, of the lesion for ISIC-2016, ISIC-2017, and ISIC-2018 datasets reveal that most of the ROIs have an aspect ratio of $1:1$. So, all the ROIs have been resized to 192×192 pixels using a nearest-neighbor interpolation for classification using the DermoExpert. The categorical cross-entropy function is used as a loss function in our DermoExpert. However, it is very impractical to guesstimate the proper optimizer and the learning rate as they are highly dependent on the networks and datasets. In this literature, we perform extensive experiments for selecting those hyperparameters as described in subsection 3.2. We set the initial epochs as 200 and stop the training using a callback function when the validation loss has stopped improving. However, we perform several extensive experiments to achieve the highest possible performance for a robust and accurate SLC system. Firstly, we select the optimizer with learning rate and the best hybrid-CNN classifier (see in subsection 2.1.3), for the proposed DermoExpert, via comprehensive experiments. Then, we perform the following experiments:

1. We fine-tuned DSNet on the ISIC-n dataset, where $n = 2016, 2017, 2018$. Then, we extract the lesion ROI and resize the images to 192×192.

2. We perform the classification using the proposed DermoExpert on those segmented ROIs.

3. We rebalance the lesion classes of the segmented ROIs since the class distributions are imbalanced. Then, we perform the classification.

4. Finally, we add the intensity and geometry-based augmentation on the experiments-3 and perform the classification.

We have repeated all the above-mentioned experiments for the three different datasets such as ISIC-2016, ISIC-2017, and ISIC-2018.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

3. Experimental Results

In this section, we are reporting the results obtained through several extensive experiments, both qualitatively and quantitatively. First, the segmentation results of the fine-tuned DSNet and the qualitative results for different augmentations are reported in subsection 3.1. Then, in subsection 3.2, we present ablation studies for optimizer and learning rate selection as they play the crucial roles for training the CNNs, and we also present the ablation studies, in subsection 3.3, for the best hybrid-CNN classifier selection (see in subsection 2.1.3) for the DermoExpert. Then, the classification results on ISIC-2016, ISIC-2017, and ISIC-2018 datasets are presented in the subsections 3.4, 3.5, and 3.6, respectively. Finally, in subsection 3.7, we compare our results with the state-of-the-art results for lesion classification.

3.1. Segmentation and Augmentation

Table 2 shows the quantitative results for the lesion ROI extraction (segmentation) for further classification. The results, as shown in Table 2, demonstrate that the type-I errors

Table 2: Segmentation results on the test datasets of the ISIC-2016, ISIC-2017, and ISIC-2018 from the fine-tuned DSNet, where the mRc, mSp, and mIoU respectively indicate the mean recall, specificity, and IoU.

Fine-tune datasets	Testing datasets	Performance metrics		
		mRc	mSp	mIoU
ISIC-2016	ISIC-2016	0.908 ± 0.12	0.962 ± 0.09	0.859 ± 0.13
	ISIC-2017	0.880 ± 0.17	0.935 ± 0.15	0.794 ± 0.17
ISIC-2018	ISIC-2018	0.911 ± 0.14	0.954 ± 0.12	0.838 ± 0.14

are 3.8 %, 6.5 %, and 4.6 % respectively for ISIC-2016, ISIC-2017, and ISIC-2018 datasets, whereas the type-II errors are 9.2 %, 12.0 %, and 8.9 % respectively for ISIC-2016, ISIC-2017, and ISIC-2018 datasets. Such a less type-I and type-II errors reveal that our segmented ROIs, of the skin lesion, are blessed with less false negative and positive respectively, which make them better-applicable for the SLC. The qualitative results of the segmented masks,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

as shown in Fig. 9, depict that the extracted green ROIs are approximately coincided on the true yellow ROIs. More segmentation results for all the test images are available in YouTube (ISIC-2016[2], ISIC-2017[3], and ISIC-2018[4]). The qualitative results, as shown in Fig. 9, show that the segmented mask is also as better as in the microscopic skin images (first and last columns of the second row). Both the quantitative and qualitative representation of the

Figure 9: The example of some extracted ROIs on the ISIC-2016, ISIC-2017, and ISIC-2018 test datasets from the fine-tuned DSNet, where the green and yellow color denote the extracted and true bounding boxes.

segmentation results demonstrate that the mean overlapping between the true and predicted mask, of all the test images, are as high as the recent state-of-the-art results for dermoscopic lesion segmentation (Bi et al., 2019; Tang et al., 2020b; Xie et al., 2020; Zhu et al., 2020). However, all the segmented ROIs are further processed for the augmentation as both the quantitative and qualitative results of the segmentation show that it yields the most reliable ROIs for the SLC.

Fig. 10 shows the typical examples of the augmented images of the segmented ROIs. The qualitative presentation of the augmented images reveals the distinctiveness of each image, which is very crucial for the training of CNNs. The image, as in Fig. 10 (d), shows that the sigmoid corrected image provides the specified lesion region to the network for learning about the lesion. The gamma-corrected ($\gamma = 1.7$), as in Fig. 10 (f), provides more intense lesion area, whereas the gamma-corrected ($\gamma = 0.7$), as in Fig. 10 (e), is brighter than the

[2]ISIC-2016 (Segmentation): `https://youtu.be/kB0Bf5D0WsA`
[3]ISIC-2017 (Segmentation): `https://youtu.be/m3u58LN9lns`
[4]ISIC-2018 (Segmentation): `https://youtu.be/r4hxv8WdQHM`

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

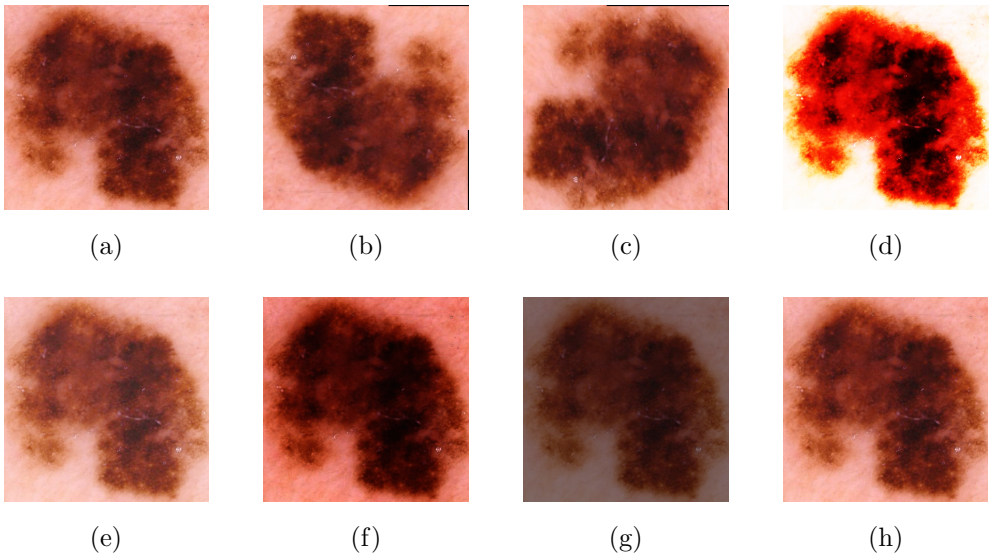


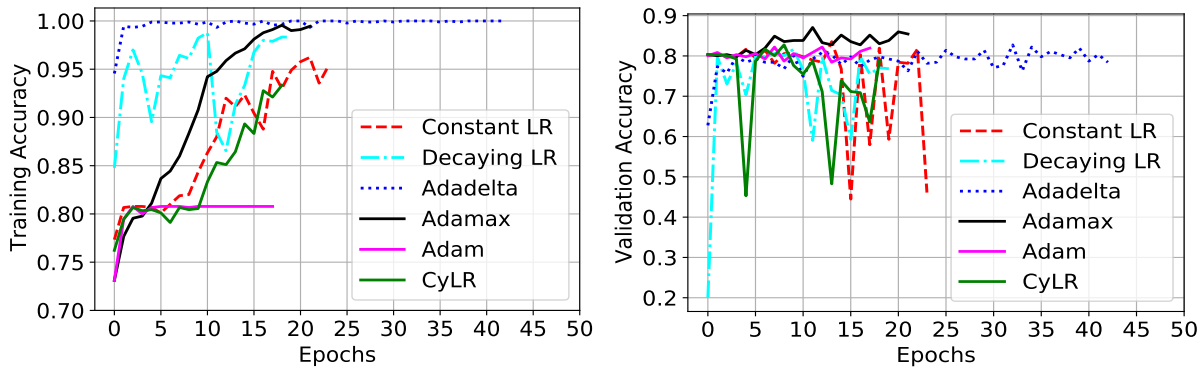
Figure 10: Example of the geometric and intensity-based augmentations of dermoscopic images, where (a) the original image, (b) 180° rotated, (c) 270° rotated, (d) sigmoid corrected, (e) gamma corrected ($\gamma = 0.7$), (f) gamma corrected ($\gamma = 1.7$), (g) logarithmic corrected, and (h) intensity re-scaling.

original image (see Fig. 10 (a)). The logarithmic corrected image, as in Fig. 10 (g), is also a darker image, which also shows the distinctiveness.

3.2. Optimizer and Learning Rate Selection

The Learning Rate (LR) is one of the most important hyperparameters, which is very challenging to select in the beginning. Too small an LR will make a training algorithm converge slowly, while too large an LR will make the training algorithm diverge (Orr and Müller, 2003). Determining a proper LR becomes more of an art than science for many machine learning problems. The optimizer, to update the weight parameters for minimizing the loss function, is also very critical and crucial to select. However, we have performed several experiments to get better-LR and optimizer for the SLC. We have used input (I_1) and preprocessing (P_1) to generate the output (O_1) by employing different optimizers and LR, which are presented in Fig. 11. This experiment has been conducted using the optimizers such as Stochastic Gradient Descent (SGD) (Ruder, 2016), Adadelta (Zeiler, 2012), Adamax & Adam (Kingma and Ba, 2014), and different LR scheduler schemes such as constant LR, decaying LR with epochs, and Cyclical Learning Rates (CyLR) (Smith, 2017). For all

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.



(a) (b)

Figure 11: The experimental results for different optimizers and LR, where (a) is for the training accuracy, and (b) is for the validation accuracy.

the experiments, we set the initial epochs of 100 with the early stopping scheme, where the training has terminated, when it stops improving the validation accuracy up to 10 epochs. The Adadelta optimizer, with an initial LR of 1.0, reaches the training accuracy of 100.0 % after few epochs, whereas the validation accuracy stagnates the improvement. Adam optimizer also has stopped, improving in terms of training and validation accuracy, with the increased epochs. The constant LR, decaying LR, and CyLR with SGD also produce many overshoots and undershoots in training and validation accuracy. On the other hand, the adaptive optimizer Adamax, with an initial LR of 0.0001, has smoothly increased training accuracy with the highest validation accuracy, where we employed LR scheduler (reduction of initial LR after 5 epochs if validation accuracy doesn't improve) along with the Adamax. However, those experiments reveal that an Adamax, with the LR scheduler, is a better choice for the SLC in our proposed DermoExpert, which is employed in the rest of the upcoming experiments.

3.3. Classifier Selection

In this subsection, we present the ablation studies on the proposed hybrid-CNN classifiers, as described in subsection 2.1.3, by comparing them quantitatively. We have used input (I_1) and preprocessing (P_1) to generate the output (O_1) for five different classifiers,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

where we also apply the best optimizer and LR from the previous experiment. The results of this experiment are presented in the ROC curve, as shown in Fig. 12, where we have also reported AUC for different classifiers. The results, in Fig. 12, show that the Method-1

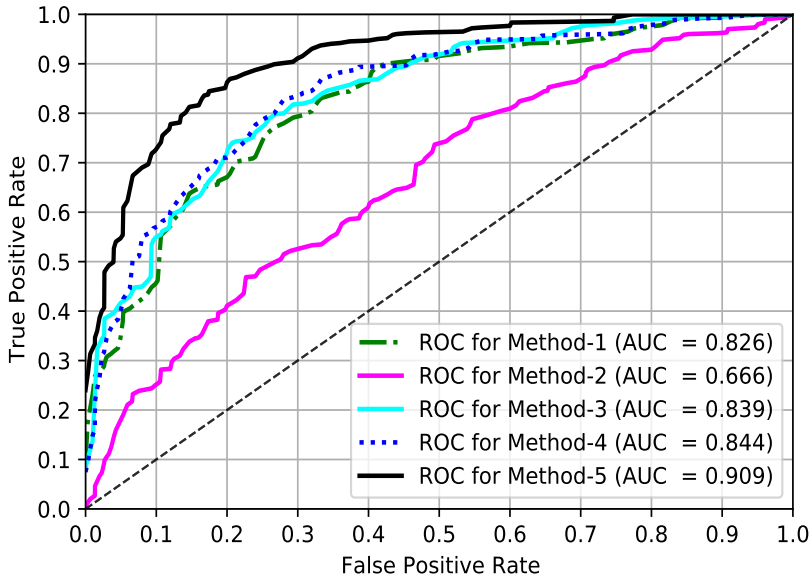


Figure 12: The ROC curve for the ISIC-2016 test dataset by employing the proposed five different networks and preprocessing (P_3).

outperforms the Method-2 by a margin of 16.0 %, in term of AUC, as well as at 10.0 % false-positive rates, the true-positive rates are approximately 50.0 % and 25.0 % respectively for Method-1 and Method-2. Those results indicate that the channel-wise concatenation (X_{FC}) is better-fusion technique than the averaging (X_{FA}) of the FMGs as the former has 3 times more depth information about the lesion. The comparison between the figures (see Fig. 8 (b), Fig. 8 (c), Fig. 8 (d), Fig. 8 (e), and Fig. 8 (f)) also demonstrates that the addition of those feature maps produces the scattered feature distribution, whereas the individual feature map from the FMG-1, FMG-2, and FMG-3 depicts better-feature maps of the lesion (see Fig. 8 (b)). Those phenomena could be the possible reasons for achieving better results from the Method-1 than Method-2. Again, Method-3, where we have employed only the second level ensembling, beats the former two methods, which reveals that second-level ensembling has good prospects in the proposed classifier than the first-level ensembling

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

alone. The employment of the fusion-by-averaging (X_{FA}) with all three distinct feature maps, called as Method-4, further improves the results of Method-3 by a margin of 0.5 % concerning the AUC. Finally, in Method-5, if we replace X_{FA} by fusion-by-concatenation (X_{FC}) in Method-4, it shows a better performance comparing all the former methods as X_{FC} has a better representation of the lesion than the X_{FA}. However, the Method-5 beats all the other four methods by the margin of 8.3 %, 24.3 %, 7.0 %, and 6.5 % concerning AUC respectively for Method-1, Method-2, Method-3, and Method-4. The above discussion reveals that the hybrid-CNN classifier (Method-5) has better-prospect for the SLC, which comprises of both the first and second-level ensembling. In the first level, we have used fusion-by-concatenation, and in the second-level, we aggregate the individual probability to get final prediction probability. However, for all the next experiments for the SLC on the ISIC-2016, ISIC-2017, and ISIC-2018 datasets, we will use our proposed hybrid-CNN classifier (Method-5) as it has better-prospects as a classifier for the robust DermoExpert.

3.4. Results on ISIC-2016

The performance, of the binary SLC of the proposed DermoExpert, has been validated using 379 dermoscopic test images of the ISIC-2016 dataset. The overall quantitative results, from all the extensive experiments, are presented, in Table 3, in terms of the recall, precision, and F1-score. The results, as presented in Table 3, show that the preprocessing

Table 3: The classification results on the ISIC-2016 test dataset from the different extensive experiments.

Preprocessing	Recall			Precision			F1-score		
	Nev	Mel	W. Avg.	Nev	Mel	W. Avg.	Nev	Mel	W. Avg.
P_1	0.99	0.35	0.87	0.86	0.93	0.87	0.92	0.50	0.84
P_2	0.98	0.69	0.93	0.93	0.91	0.93	0.96	0.79	0.92
P_3	0.93	0.85	0.92	0.96	0.76	0.92	0.95	0.81	0.92

P_1: Segmentation; P_2: Segmentation and Rebalancing; P_3: Segmentation, Rebalancing, and Augmentation.

(P_3) along with the proposed hybrid-CNN classifier yields the best performance for binary lesion classification. The recall of the positive class (Mel) reveals that the type-II errors are

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

65.0 %, 31.0 %, and 15.0 % for the respective preprocessing P_1, P_2, and P_3. It also shows that rebalancing with segmentation improves the type-II errors by 34.0 %, whereas the rebalancing and augmentation with segmentation significantly improves the type-II errors by 50.0 %. Although the FN is reducing, when we move from P_1 to P_3, the FP is increasing accordingly. For a P_3 preprocessing, the precision of the Mel class (76.0 %) also shows the evidence of increasing FP with the decreasing of FN, as 24.0 % recognized positives are the wrong positives. However, the decreasing FN-rates (65.0 % to 15.0 %) is better than the increasing FP-rates (1.0 % to 7.0 %) in medical diagnosis applications. Moreover, the harmonic mean of the precision and recall for both the classes (Nev and Mel) is improving, when we move from P_1 to P_3 by the margins of 3.0 % and 31.0 % respectively for Nev and Mel classes. The confusion matrix in Table 4, for more detailed analysis of the SLC-2016

Table 4: The confusion matrix for the ISIC-2016 test dataset by using the proposed CNN-based hybrid classifier and preprocessing (P_3).

		Actual	
		Nev	Mel
Predicted	Nev	284 93.42 %	11 14.67 %
	Mel	20 6.58 %	64 85.33 %

results, shows that among 304 Nev samples, correctly classified samples are 284 (93.4 %), whereas only 20 (6.58 %) samples are classified as Mel. It also shows that among 75 Mel samples, correctly classified samples are 64 (85.33 %), whereas only 11 (14.67 %) samples are classified as Nev. Fig. 13 shows the ROC curve of the best SLC-2016 and the baseline Xception (Chollet, 2017), ResNet-50 (He et al., 2016a), and DenseNet-121 (Huang et al., 2017). The proposed DermoExpert obtains an AUC of 0.96, which indicates the probability of correct lesion recognition is as high as 96.0 % for any given random sample. It has beaten the baseline Xception, ResNet-50, and DenseNet-121 respectively by 11.6 %, 13.0 %, and 10.6 % in terms of AUC. Also from Fig. 13 and given a 10.0 % false-positive rates, the

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

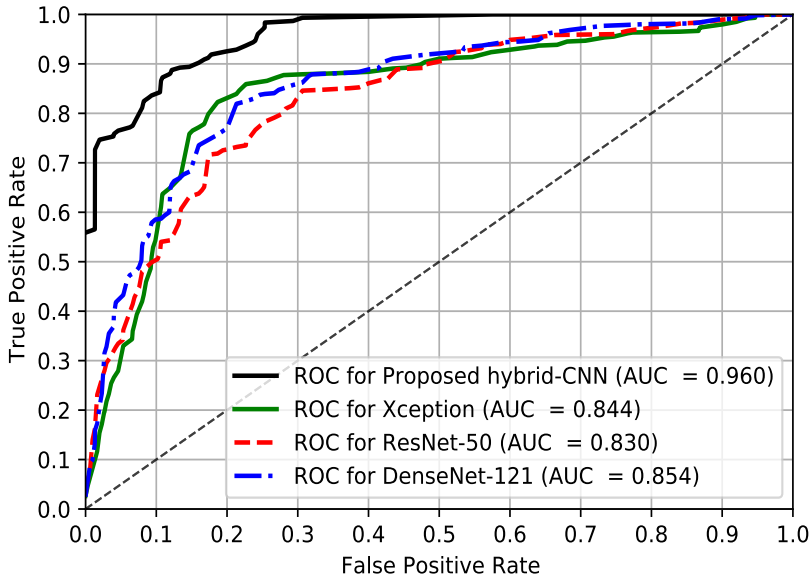


Figure 13: The ROC curve for the ISIC-2016 test dataset by employing the proposed CNN-based hybrid classifier and preprocessing (P_3).

true-positive rates of the proposed DermoExpert, Xception, ResNet-50, and DenseNet-121 are approximately 83.0 %, 55.0 %, 50.0 %, and 58.0 % respectively. The above-discussions, for the SLC on ISIC-2016 test dataset, indicate that the proposed DermoExpert can be potentially used as a CAD tool for the lesion classification.

*3.5. Results on ISIC-*2017

This subsection represents the potentiality of the proposed DermoExpert to recognize three different lesions as Nev, SK, and Mel. The quantitative results, on the ISIC-2017 test dataset, have been summarized, in Table 5, for the recall, precision, and F1-score. The results, as in Table 5, also depicts the weighted metrics for aggregate cases concerning the class population. The SLC results, as shown in Table 5, demonstrate that the recall is increased by 24.0 %, and 17.0 % respectively for Mel and SK, when we employ the preprocess P_2 instead of baseline P_1. The further employment of the preprocess P_3 in place of P_2 could not reduce the type-II errors rather remains constant, but reduces type-II errors of the Nev class by 5.0 %. The weighted average of the recall has increased, when we employ the preprocess P_2 instead of baseline P_1, and then the preprocess P_3 instead of baseline P_2.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Table 5: The classification results on the ISIC-2017 test dataset from the different extensive experiments.

Preprocessing	Recall				Precision				F1-score			
	Nev	SK	Mel	W. Avg.	Nev	SK	Mel	W. Avg.	Nev	SK	Mel	W. Avg.
P1	0.98	0.67	0.38	0.81	0.80	0.92	0.80	0.82	0.88	0.77	0.51	0.79
P2	0.90	0.84	0.62	0.84	0.88	0.68	0.84	0.84	0.89	0.75	0.71	0.84
P3	0.95	0.84	0.61	0.86	0.90	0.76	0.82	0.86	0.92	0.80	0.70	0.86

P_1: Segmentation; P_2: Segmentation and Rebalancing; P_3: Segmentation, Rebalancing, and Augmentation.

However, it is beneficiary of applying the preprocess P_3 instead of P_2 and P_1 in terms of type-II errors. It is noticed, from Table 5, that the weighted precisions are increasing, when we change the preprocess P_1 to P_2 and P_2 to P_3. The class-wise precision also reveals that the FN is reducing by a significant amount, although the FP is increasing. However, such a result in medical disease diagnosis, for the SLC, is acceptable as the patient, with the positive symptom, should not be classified as the negative patient. Additionally, the class-wise F1-score for the Nev, SK, and Mel have improved significantly, when we change the prepossess P_1 to P_2, and then P_2 to P_3. The improved F1-score tells that both the recall and precision are praiseworthy, although the uneven class distribution was being used. Details of class-wise investigation, of the best performing SLC, has been present in the confusion matrix in Table 6, where we apply the preprocessing (P_3) and the proposed hybrid-CNN classifier. The confusion matrix, as presented in Table 6, for the SLC-2017 shows that

Table 6: The confusion matrix for the ISIC-2017 test dataset by using the proposed hybrid-CNN classifier and preprocessing (P_3).

		Actual		
		Nev	SK	Mel
Predicted	Nev	372 94.66 %	6 6.67 %	35 29.91 %
	SK	13 3.31 %	76 84.44 %	11 9.40 %
	Mel	8 2.03 %	8 8.89 %	71 60.69 %

94.66 % Nev samples are correctly classified as Nev, whereas 5.34 % (3.31 % as SK and

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

2.03 % as Mel) samples are wrongly classified. 84.44 % SK samples are correctly classified as SK, whereas 15.56 % (6.67 % as Nev and 8.89 % as Mel) samples are wrongly classified. On the other hand, 60.69 % Mel samples are correctly classified as Mel, whereas 39.31 % (29.91 % as Nev and 9.40 % as SK) samples are wrongly classified. Although the 39.31 % of the positive samples (Mel) are wrongly classified, it is still better than the baseline 62.0 % errors in the baseline preprocessing P_1. Fig. 14 shows the ROC curve of the best SLC-

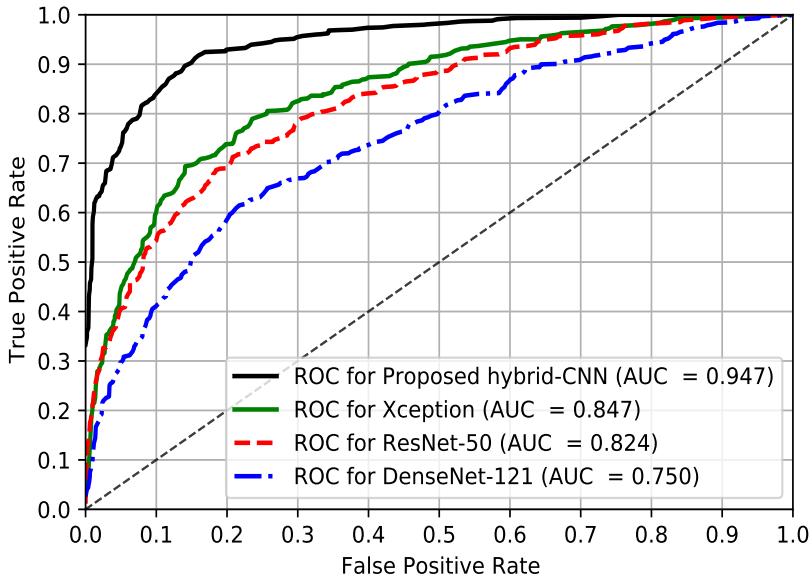


Figure 14: The ROC curve for the ISIC-2017 test dataset by employing the proposed classifier and preprocessing (P_3).

2017 and baseline state-of-the-art Xception, ResNet-50, and DenseNet-121. The proposed DermoExpert, for the SLC-2017, obtains an AUC of 0.947, which indicates the probability of correct lesion recognition is as high as 94.7 % for any given random sample. The proposed DermoExpert outperforms the Xception, ResNet-50, and DenseNet-121 respectively by 10 %, 12.3 %, and 19.7 % for AUC. Also from Fig. 14 and given a 10.0 % false-positive rates, the true-positive rates of the proposed DermoExpert, Xception, ResNet-50, and DenseNet-121 are approximately 85.0 %, 60.0 %, 55.0 %, and 41.0 % respectively. In contrast, the above-discussions for the SLC on ISIC-2017 demonstrate that the proposed DermoExpert can be potentially applied as an SLC-CAD tool.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

3.6. Results on ISIC-2018

This subsection represents the experimental results for the SLC-2018 from the proposed DermoExpert for recognizing very challenging lesion categories into seven classes such as Nev, SK, BCC, AK, DF, VL, and Mel. As it was presented earlier, in subsection 2.1.1, that the ISIC-2018 provides only the training set. Therefore, we employed 5-fold cross-validation, where 60.0 %, 20.0 %, and 20.0 % samples are used respectively for training, validation, and testing. We repeat the experiments, for the SLC-2018, 5-times and the final results, as presented in Fig. 15, are the average classification performance of the 5 folds. The investigation on the results, as shown in Fig. 15, show that for a baseline preprocessing

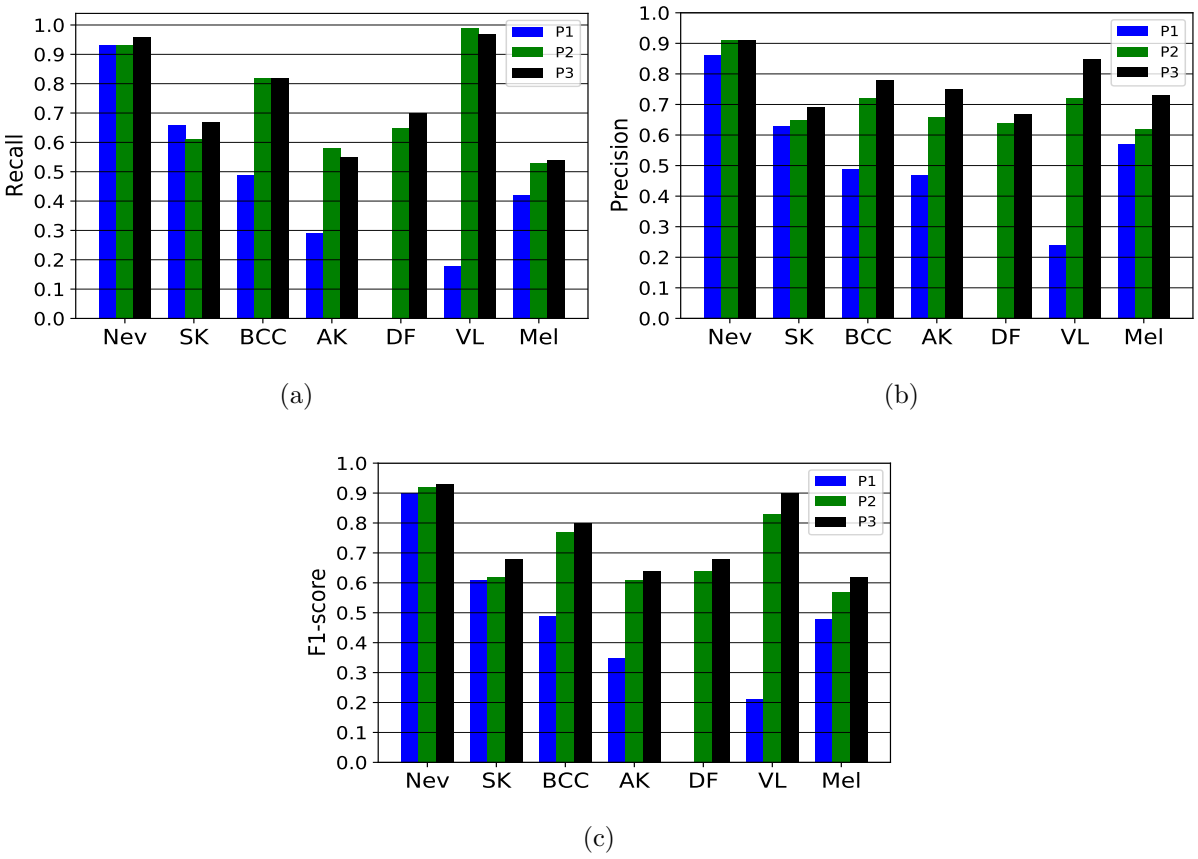


Figure 15: The experimental results for the ISIC-2018 test dataset classification, where blue, green, and black bar of different class respectively denote the results (recall (a), precision (b), and F1-score (c)) for the preprocessing P_1, P_2, and P_3.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

(P_1), the metrics (see blue bar in Fig. 15 (a), Fig. 15 (b), and Fig. 15 (c)) are varied in a significant amount, where the recall, precision, and F1-score of the Nev class are higher than other classes. Moreover, the melanoma class has a recall of 42.0 %, which is very less. Additionally, the recall, precision, and F1-score for the DF class are 0.0 %, where all the images, in the DF class, are classified as other classes. Such a weak result from the proposed DermoExpert by applying the baseline preprocessing (P_1) is due to the unequal class distribution, where class distribution was 1.0 : 6.1 : 13.02 : 20.47 : 58.3 : 47.47 : 6.02 respectively for the classes Nev, SK, BCC, AK, DF, VL, and Mel. The weak performance of the DF and VL classes, in terms of recall, precision, and F1-score, was likely to happen as they are the most minority (underrepresented) classes. However, the rebalancing the class distribution, as in the preprocessing (P_2), boosts the class-wise performance in a significant portions (see green bar in Fig. 15 (a), Fig. 15 (b), and Fig. 15 (c)), where the recall, precision, and F1-score of the DF class respectively increase to 65.0 %, 64.0 %, and 64.0 % from the baseline 0.0 for all metrics. Not only the DF class but also the other classes have improved the performance, especially the recall of BCC, AK, and Mel class has increased by the margin of 33.0 %, 29.0 %, and 11.0 % respectively. Another point can claim that the moving the preprocessing, from P_1 to P_2, do not degrade the performance of the best performing Nev class in P_1, rather improves in terms of precision and F1-score, while the recall remains constant. However, further addition of the proposed augmentation with the preprocessing (P_2) improves all the metrics for most of the classes, while the performance in the other classes remain constant. The more improved class-wise precision and F1-score, due to the appliance of the third preprocessing (P_3), show that the positive predictive value and the balanced precision-recall have more improved than the type-II errors (recall). The more detailed class-wise assessment, of the best performing SLC for the ISIC-2018 test dataset, has been presented in the confusion matrix in Table 7, where we apply the preprocessing (P_3) and the proposed hybrid-CNN classifier. The matrix, as shown in Table 7, reveals the FN and FP for the SLC-2018, where number of wrongly classified images (type-I or type-II errors) are 58/1341 (4.33 %), 72/220 (32.73 %), 18/102 (17.65 %), 29/65 (44.62 %), 7/23 (30.43 %), 1/29 (3.45 %), and 102/222 (45.94 %) respectively for the Nev, SK, BCC, AK, DF, VL and

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Table 7: The confusion matrix for the ISIC-2018 test dataset by using the proposed classifier and preprocessing (P_3).

		Actual						
		Nev	SK	BCC	AK	DF	VL	Mel
Predicted	Nev	1283 95.67 %	43 19.55 %	10 9.80 %	3 4.62 %	4 17.39 %	1 3.45 %	64 28.83 %
	SK	21 1.57 %	148 67.27 %	7 6.86 %	11 16.92 %	2 8.69 %	0 0.0 %	27 12.16 %
	BCC	9 0.67 %	4 1.82 %	84 82.36 %	8 12.31 %	0 0.0 %	0 0.0 %	3 1.35 %
	AK	0 0.0 %	9 4.09 %	0 0.0 %	36 55.38 %	0 0.0 %	0 0.0 %	3 1.35 %
	DF	5 0.37 %	2 0.91 %	0 0.0 %	1 1.54 %	16 69.57 %	0 0.0 %	0 0.0 %
	VL	0 0.0 %	0 0.0 %	0 0.0 %	0 0.0 %	0 0.0 %	28 96.55 %	5 2.25 %
	Mel	23 1.72 %	14 6.36 %	1 0.98 %	6 9.23 %	1 4.35 %	0 0.0 %	120 54.06 %

Mel. Although the performance, in some classes, is not highly improved as in other classes it is still better than the baseline process. Fig. 16 shows the ROC curve of the best SLC-2018 and baseline state-of-the-art Xception, ResNet-50, and DenseNet-121. The proposed DermoExpert, for SLC-2018, obtains an AUC of 0.969, which indicates the probability of correct lesion recognition is as high as 96.9 % for any given random sample. It has defeated all the baseline Xception, ResNet-50, and DenseNet-121 respectively by 2.4 %, 2.4 %, and 2.3 % with respect to AUC. Also from Fig. 16 and given a 10 % false-positive rates, the true-positive rates of the proposed DermoExpert, Xception, ResNet-50, and DenseNet-121 are approximately 92.0 %, 83.0 %, 82.0 %, and 86.0 % respectively. In conclusion, the above-discussions for the SLC on ISIC-2018 demonstrate that the proposed DermoExpert can be potentially applied as a SLC-CAD tool.

3.7. Results Comparison

In this subsection, the performance of the proposed DermoExpert is compared and contrasted to several recent state-of-the-art methods.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

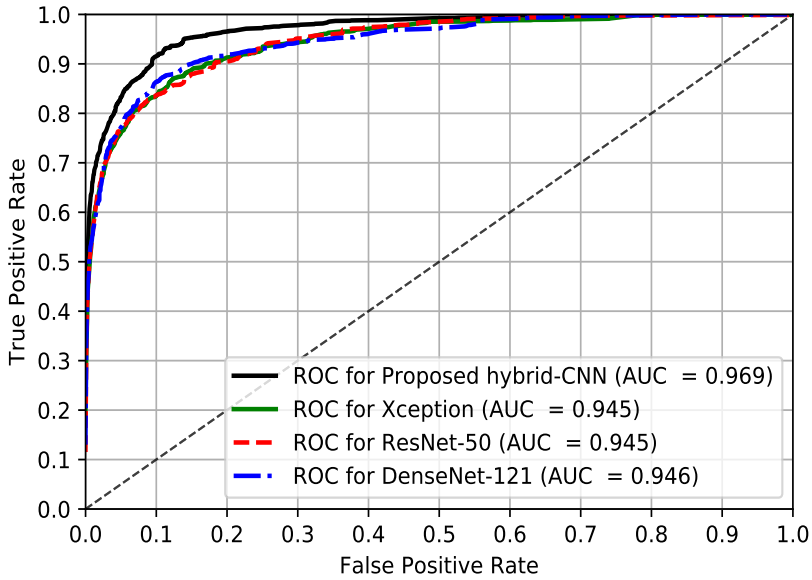


Figure 16: The ROC curve for the ISIC-2018 test dataset by employing the proposed network and preprocessing (P_3).

Table 8 represents the performance comparison of the proposed DermoExpert with other recent methods for all the ISIC-2016, ISIC-2017, and ISIC-2018 datasets. To improve classification performance, authors, in several recent works, used the external data to train their models, which are not publicly available yet. The improvement of the classification network may not be due to the superiority of the network itself, but the characteristics of the external data, which are similar to the test datasets. However, for the fairness in comparison, we have reported the results, which were too strict the datasets on the ISIC archive only. The proposed DermoExpert produces the best classification, as shown in Table 8, for five out of the nine cases while performing second best with the winning methods on the other four cases.

Comparison of SLC-2016**.** The proposed DermoExpert produces the best results for the AUC by beating the state-of-the-art (Tang et al., 2020a; Yu et al., 2020) with a 10.0 % margin. Concerning the type-II errors (recall), DermoExpert is behind the state-of-the-art (Song et al., 2020) by 7.0 %, but the DermoExpert outperforms the FPRPN (Song et al., 2020) by a 10.0 % margin concerning the positive predictive value (precision). However, in

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Table 8: The state-of-the-art comparison with proposed DermoExpert, which were trained, validated and tested on the ISIC-2016, ISIC-2017, and ISIC-2018 datasets.

Methods	ISIC-2016			ISIC-2017			ISIC-2018		
	Recall	Precision	AUC	Recall	Precision	AUC	Recall	Precision	AUC
ResNet-50 (Brinker et al., 2019)	0.56	0.71	0.85	-	-	-	-	-	-
GR (Serte and Demirel, 2019)	-	-	-	0.15	-	0.91	-	-	-
ARLCNN (Zhang et al., 2019)	-	-	-	0.77	-	0.92	-	-	-
IR (Al-Masni et al., 2020)	0.82	-	0.77	0.76	-	-	0.81	-	-
FPRPN (Song et al., 2020)	**0.99**	0.82	0.81	**0.98**	0.82	0.79	-	-	-
GCD (Tang et al., 2020a)	0.32	0.73	0.86	0.55	-	0.93	-	-	-
MFA (Yu et al., 2020)	0.60	0.69	0.86	-	0.72	0.90	-	-	
MIAG (MIA Group, 2018)	-	-	-	-	-	-	0.78	**0.87**	**0.98**
DenseNet (Li and Li, 2018)	-	-	-	-	-	-	0.80	0.85	**0.98**
RDI (Nozdryn-Plotnicki et al., 2018)	-	-	-	-	-	-	0.83	0.83	**0.98**
DermoExpert (Proposed, 2020)	0.92	**0.92**	**0.96**	0.86	**0.86**	**0.95**	**0.86**	0.85	0.97

GR: Gabor Wavelet-based CNN (Serte and Demirel, 2019)
ARLCNN: Attention residual learning convolutional neural network (ResNet-14 & ResNet-50) (Zhang et al., 2020)
IR: Inception-ResNet-V2 (ISIC-2016), ResNet-50 (ISIC-2017,ISIC-2018) (Al-Masni et al., 2020)
FPRPN: Feature Pyramid Network (FPN) and Region Proposal Network (RPN) (Song et al., 2020)
GCD: Global-Part CNN Model with Data-Transformed Ensemble Learning (Tang et al., 2020a)
MFA: Multi-network based feature aggregation (Yu et al., 2020)
MIAG: Emsembling classifier (MIA Group, 2018)
RDI: Emsembling of ResNet, DenseNet, and Inception (Nozdryn-Plotnicki et al., 2018)

terms of balanced accuracy (avg. of recall and precision), DermoExpert beats the state-of-the-art (Song et al., 2020) by a 1.5 % margin.

Comparison of SLC-2017**.** For AUC, the DermoExpert wins by defeating the second-highest (Tang et al., 2020a) by a margin of 2.0 %. DermoExpert has beaten the work of Song et al. (2020) by the margins of 4.0 %, and 16.0 % with respective precision and AUC although it lost by Song et al. (2020) in terms of recall. However, the proposed DermoExpert produces the second-highest results by beating the third results (Zhang et al., 2019) with a margin of 9.0 % for recall.

Comparison of SLC-2018**.** The results of DermoExpert for ISIC-2018 are compared with the top three performers of the ISIC-2018 competition leaderboard (Li and Li, 2018; MIA Group, 2018; Nozdryn-Plotnicki et al., 2018), where DermoExpert wins in recall and serves as a second winner in other two metrics. It also beats the recently published method of Al-Masni et al. (2020) by a margin of 5.0 % in terms of type-II errors. It means that 5.0 % additional samples will be classified correctly as the true class than the method of Al-Masni et al. (2020). However, in terms of the balanced accuracy, DermoExpert beats the challenge

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

topper by a margin of 3.0 %. The recent method proposed by Mahbod et al. (2020), for the test results of ISIC-2018, are not presented in Table 8, as they did not use the same dataset to train their model, as in the proposed DermoExpert.

4. Discussion and Application

CNN-based classifiers are better-choice in different medical imaging context, where they automatically learn low-, middle-, and high-level features directly from the input images. Finally, fully connected neural networks, also known as multilayer perceptron, classify those features. However, training of such deep CNN-based classifiers is an arduously challenging process, especially when the train with a smaller dataset as in the ISIC skin lesion datasets. There are two commonly occurring limitations in current CNN-based classifiers, namely: prone to overfitting and vanishing gradient problem. Those two crucial limitations are reduced, in this research, for the SLC.

To build a generic and robust end-to-end SLC system utilizing a smaller dataset, we proposed a hybrid classifier, where we use three distinct feature map generators rather than single generator as in Xception, ResNet-50, DenseNet-121, and *etc.* In each generator in the proposed classifier, we use several skip connections, which enable each layer of the generator to directly access to the gradients of all previous layers. However, such a design, of the CNN network, alleviates the vanishing gradient problem, and partly reduces the overfitting as the final feature map has a deeper presentation of the skin lesion. The adaptation of pre-trained weights in place of the training of deep CNN classifier from scratch, with any random initialization, also party reduces the overfitting and improves the performance. The experimental results, as in subsection 3.3, have validated that multiple features from different generators provide better results. Moreover, channel-wise concatenation of different features is better than the addition of them as the latter one generates scattered feature maps (see Fig. 8). In the same experimental conditions, the former approach outperforms the latter approach by a margin of 16.0 % concerning AUC (see Fig. 12). Further addition of second-level ensembling, as described in subsection 2.1.3, provides 7.0 % more AUC than first- or second-level ensembling alone (see Fig. 12). Instead of a random selection of LR

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

and optimizer, it is better to conduct several extensive experiments, as the cost functions depend on the data distribution, size, and inter- or intra-class variability. However, our experimental results, as presented in subsection 3.2, have demonstrated that the adaptive optimizer Adamax, with an initial LR of 0.0001, has better-prospect for the SLC (see Fig. 11) when the LR scheduler is incorporated with the Adamax.

The classification results by the proposed DermoExpert, for the ISIC-2016 (see subsection 3.4), ISIC-2017 (see subsection 3.5), and ISIC-2018 (see subsection 3.6), show that class rebalancing along with the segmentation boost the performance of underrepresented class. The F1-score of the test datasets of ISIC-2016, ISIC-2017, and ISIC-2018 are improved by the margin of 8.0 %, 5.0 %, and 1.0 %, respectively, when we have added extra images to underrepresented class and weighted the loss function. Further addition of image augmentation, with segmentation and rebalancing, improves the classification performance. It also reduces the overfitting, for the lesion classification, by reducing the differences between the training and testing performances. For all the test datasets of lesion classification, the proposed DermoExpert, with multiple feature map generators and two-level ensembling, performs better than the classifier having a single generator (see Fig. 13, Fig. 14, and Fig. 16). The experimental results, in subsections 3.4, 3.5, and 3.6, also demonstrated that the classification performance for ISIC-2016 (two classes) is better than ISIC-2017 (three classes). The addition of SK class, as in ISIC-2017, reduces the F1-score of Nev and Mel classes respectively by the margin of 3.0 %, and 11.0 %. The higher similarity of SK with Nev and Mel classes is the possibility of such reduced performance in ISIC-2017 test results. Moreover, a more number of classes tend to bring complications in the classifiers, especially when the training with fewer examples and intra-class similarity, as in the SLC on the ISIC skin lesion dataset. It is also observable that the classification performance, for ISIC-2018, has improved, although it has more number of classes (7 classes). However, the data distribution of all ISIC, as shown in Table 1, depicts that ISIC-2018 has much higher samples than ISIC-2016 and ISIC-2017.

However, those discussion reveals the superiority of the proposed DermoExpert for the skin lesion classification, which shows it's acceptance for the SLC-CAD system. A few qualitative results, of the proposed DermoExpert, are illustrated in Fig. 17, where the seg-

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

mented masks from the fine-tuned DSNet are used to detect the lesion ROI (yellow color contour) along with the recognized class. More classification results for all the test images are available in YouTube (ISIC-2016[5], ISIC-2017[6], and ISIC-2018[7]). Fig. 17 also shows a

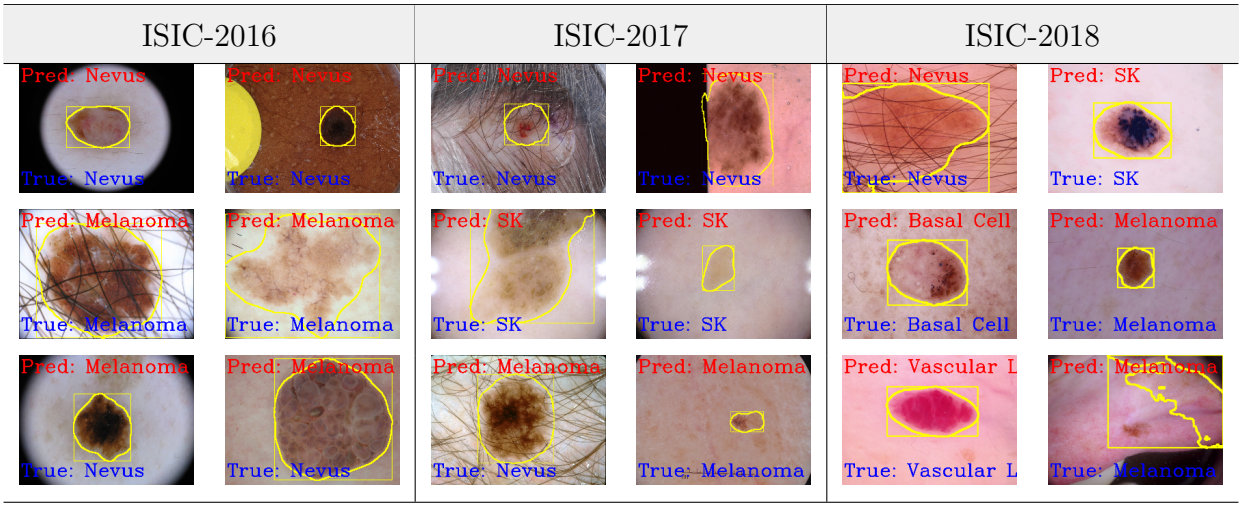
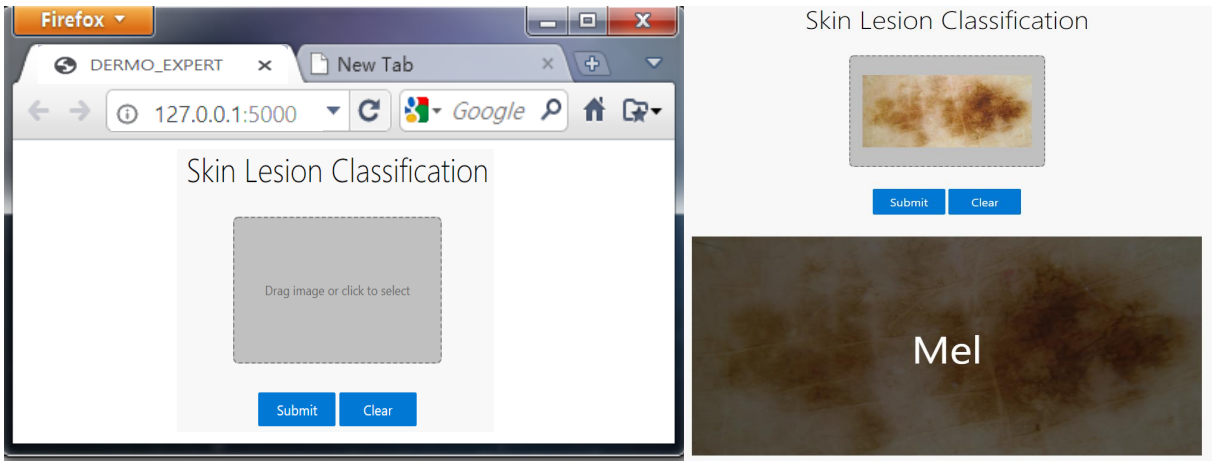


Figure 17: The qualitative classification results using the DermoExpert, where the classification has been accomplished using the segmented ROIs (yellow color) and the proposed DermoExpert.

few challenging images to be classified as well as some wrongly recognized images. Although those images are wrongly predicted by the DermoExpert, they visually seem like the predicted class. However, in this article, we have presented the prospective applications of the DermoExpert by building a web application, as shown in Fig. 18 (a), for the user of the DermoExpert. We have implemented the app in our local machine, which runs in a web browser at "http://127.0.0.1:5000/" by accessing the CNN environments of that local machine. Our source code, and segmented masks, for ISIC-2018 dataset, will be made publicly available in GitHub[8].

[5]ISIC-2016 (Classification): `https://youtu.be/wwHwkQmigqU`
[6]ISIC-2017 (Classification): `https://youtu.be/1Dn5l4g4h6Y`
[7]ISIC-2018 (Classification): `https://youtu.be/NXVw2cyqd6k`
[8]`https://github.com/kamruleee51/Web-App-of-Skin-Lesion-Classification`

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.



(a) (b)

Figure 18: The web app, for the prediction of the lesion class, by employing the trained weights of the DermoExpert. Users can drag or choose an image by clicking the button, as shown on the left side, as well as they can clear if the selection is wrong. After that, by clicking the submit button, the user can see the recognized result on the screen, as shown on the right side.

5. Conclusion

Automatic recognition of skin lesions is very important, although it is very challenging due to high visual similarity and diverse artifacts. However, in this article, the skin lesion recognition has been automated by proposing a pipeline called DermoExpert. This article emphasized on a systematic evaluation of an integrated skin lesion recognition system including lesion ROI extraction, image augmentation, rebalancing, and hybrid-CNN classifier. Our experimental results demonstrate that the proposed DermoExpert can discriminate against the lesion features more accurately as we concatenated features from three distinct generators. Thus, it achieves state-of-the-art performance for the SLC of three different datasets. The segmented skin lesions, rather than the whole images, can provide more salient and representative features from the CNNs, which can lead to the improvement of the SLC. Moreover, the rebalanced the class distribution attained better performance of the SLC as compared to the imbalanced distribution. Additionally, the augmentation can lead the CNN-based classifier to be more generic as CNNs can learn from diverse training

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

samples. In the future, we will further explore and investigate the effects of the improvement of segmentation and weighting of the underrepresented classes. The weights of the DermoExpert will be deployed to the Google Cloud platform to make it publicly available. We will provide our segmented masks of the ISIC-2018 dataset (Task 3: Lesion Diagnosis ($10,015$ images)) for the research purpose (on-request) as they are not available yet. The proposed DermoExpert will be applied to other domains for medical imaging to verify its versatility and generality.

Acknowledgements

None. No funding to declare.

Conflict of Interest

All authors have no conflict of interest to publish this research.

References

Ahonen, T., Hadid, A., Pietikainen, M., 2006. Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence 28, 2037–2041.

Akar, E., Marques, O., Andrews, W., Furht, B., 2019. Cloud-based skin lesion diagnosis system using convolutional neural networks, in: Intelligent Computing-Proceedings of the Computing Conference, Springer. pp. 982–1000.

Al-Masni, M.A., Kim, D.H., Kim, T.S., 2020. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Computer Methods and Programs in Biomedicine 190, 105351.

Almaraz-Damian, J.A., Ponomaryov, V., Sadovnychiy, S., Castillejos-Fernandez, H., 2020. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. Entropy 22, 484.

American Institute for Cancer Research, 2018. *Skin cancer statistics*. `https://www.wcrf.org/dietandcancer/cancer-trends/skin-cancer-statistics` [Accessed: 08 May 2020].

Amin, J., Sharif, A., Gul, N., Anjum, M.A., Nisar, M.W., Azam, F., Bukhari, S.A.C., 2020. Integrated design of deep features fusion for localization and classification of skin cancer. Pattern Recognition Letters 131, 63–70.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Beale, H.D., Demuth, H.B., Hagan, M., 1996. Neural network design. Pws, Boston .

Bi, L., Kim, J., Ahn, E., Kumar, A., Feng, D., Fulham, M., 2019. Step-wise integration of deep class-specific learning for dermoscopic image segmentation. Pattern recognition 85, 78–89.

Brinker, T.J., Hekler, A., Enk, A.H., von Kalle, C., 2019. Enhanced classifier training to improve precision of a convolutional neural network to identify images of skin lesions. PloS one 14.

Brinker, T.J., Hekler, A., Utikal, J.S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A.H., von Kalle, C., 2018. Skin cancer classification using convolutional neural networks: systematic review. Journal of medical Internet research 20, e11936.

Chaturvedi, S.S., Gupta, K., Prasad, P.S., 2020. Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using mobilenet, in: International Conference on Advanced Machine Learning Technologies and Applications, Springer. pp. 165–176.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al., 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv:1902.03368 .

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 168–172.

Cunningham, P., Delany, S.J., 2007. k-nearest neighbour classifiers. Multiple Classifier Systems 34, 1–17.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Friedman, R.J., Rigel, D.S., Kopf, A.W., 1985. Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. CA: a cancer journal for clinicians 35, 130–151.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906–914.

Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., Garnavi, R., 2017. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 250–258.

Géron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A., 2020. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. MethodsX , 100864.

Ghalejoogh, G.S., Kordy, H.M., Ebrahimi, F., 2020. A hierarchical structure based on stacking approach for skin lesion classification. Expert Systems with Applications 145, 113127.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 249–256.

Guo, S., Yang, Z., 2018. Multi-channel-resnet: An integration framework towards skin lesion analysis. Informatics in Medicine Unlocked 12, 67–74.

Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A., 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv:1605.01397 .

Hameed, N., Shabut, A.M., Ghosh, M.K., Hossain, M., 2020. Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. Expert Systems with Applications 141, 112961.

Harangi, B., 2018. Skin lesion classification with ensembles of deep convolutional neural networks. Journal of biomedical informatics 86, 25–32.

Hasan, M.K., Dahal, L., Samarakoon, P.N., Tushar, F.I., Martí, R., 2020. DSNet: Automatic dermoscopic skin lesion segmentation. Computers in Biology and Medicine 120, 103738.

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

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Identity mappings in deep residual networks, in: European conference on computer vision, Springer. pp. 630–645.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: Achievements and challenges. Journal of digital imaging 32, 582–596.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 .

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

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

Hussain, Z., Gimenez, F., Yi, D., Rubin, D., 2017. Differential data augmentation techniques for medical imaging classification tasks, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 979.

ISIC, 2018. *ISIC Archive*. `https://www.isic-archive.com/\#!/topWithHeader/onlyHeaderTop/`

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

gallery [Accessed: 09 May 2020].

Jalalian, A., Mashohor, S., Mahmud, R., Karasfi, B., Saripan, M.I.B., Ramli, A.R.B., 2017. Foundation and methodologies in computer-aided diagnosis systems for breast cancer detection. EXCLI Journal 16, 113–137.

Jones, O., Jurascheck, L., van Melle, M., Hickman, S., Burrows, N., Hall, P., Emery, J., Walter, F., 2019. Dermoscopy for melanoma detection and triage in primary care: a systematic review. BMJ open 9, e027529.

Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172, 1122–1131.

Khan, M.A., Sharif, M., Akram, T., Bukhari, S.A.C., Nayak, R.S., 2020. Developed newton-raphson based deep features selection framework for skin lesion recognition. Pattern Recognition Letters 129, 293–303.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980 .

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D., 2016. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE journal of biomedical and health informatics 21, 31–40.

Kwasigroch, A., Grochowski, M., Mikołajczyk, A., 2020. Neural architecture search for skin lesion classification. IEEE Access 8, 9061–9071.

Li, K.M., Li, E.C., 2018. Skin lesion analysis towards melanoma detection via end-to-end deep learning of convolutional neural networks. arXiv:1807.08332 .

Li, P., Samorodnitsk, G., Hopcroft, J., 2013. Sign cauchy projections and chi-square kernel, in: Advances in Neural Information Processing Systems, pp. 2571–2579.

Li, Y., Shen, L., 2018. Skin lesion analysis towards melanoma detection using deep learning network. Sensors 18, 556.

Liaw, A., Wiener, M., et al., 2002. Classification and regression by randomforest. R news 2, 18–22.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv:1312.4400 .

Lopez, A.R., Giro-i Nieto, X., Burdick, J., Marques, O., 2017. Skin lesion classification from dermoscopic images using deep learning techniques, in: 2017 13th IASTED international conference on biomedical engineering (BioMed), IEEE. pp. 49–54.

Ma, Z., Tavares, J.M.R., et al., 2015. A review of the quantification and classification of pigmented skin lesions: from dedicated to hand-held devices. Journal of medical systems 39, 177.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

2018. Exploring the limits of weakly supervised pretraining, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C., 2019. Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics 71, 19–29.

Mahbod, A., Schaefer, G., Wang, C., Dorffner, G., Ecker, R., Ellinger, I., 2020. Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. Computer Methods and Programs in Biomedicine , 105475.

Majtner, T., Bajić, B., Yildirim, S., Hardeberg, J.Y., Lindblad, J., Sladoje, N., 2018. Ensemble of convolutional neural networks for dermoscopic images classification. arXiv:1808.05071 .

Majtner, T., Yildirim-Yayilgan, S., Hardeberg, J.Y., 2016a. Combining deep learning and hand-crafted features for skin lesion classification, in: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE. pp. 1–6.

Majtner, T., Yildirim-Yayilgan, S., Hardeberg, J.Y., 2016b. Efficient melanoma detection using texture-based rsurf features, in: International Conference on Image Analysis and Recognition, Springer. pp. 30–37.

MIA Group, 2018. *Medical Image Analysis Group.* `https://challenge2018.isic-archive.com/leaderboards/` [Accessed: 01 April 2020].

Mishraa, N.K., Celebi, M.E., 2016. An overview of melanoma detection in dermoscopy images using image processing and machine learning. arXiv:1601.07843 .

Mporas, I., Perikos, I., Paraskevas, M., 2020. Color models for skin lesion classification from dermatoscopic images, in: Advances in Integrations of Intelligent Methods. Springer, pp. 85–98.

Narayanamurthy, V., Padmapriya, P., Noorasafrin, A., Pooja, B., Hema, K., Nithyakalyani, K., Samsuri, F., et al., 2018. Skin cancer detection using non-invasive techniques. RSC advances 8, 28095–28130.

Nozdryn-Plotnicki, A., Yap, J., Yolland, W., 2018. Ensembling convolutional neural networks for skin cancer classification. International Skin Imaging Collaboration (ISIC) Challenge on Skin Image Analysis for Melanoma Detection. MICCAI .

Orr, G.B., Müller, K.R., 2003. Neural networks: tricks of the trade. Springer.

Pereira, P.M., Fonseca-Pinto, R., Paiva, R.P., Assuncao, P.A., Tavora, L.M., Thomaz, L.A., Faria, S.M., 2020. Skin lesion classification enhancement using border-line features–the melanoma vs nevus problem. Biomedical Signal Processing and Control 57, 101765.

Pham, T.C., Luong, C.M., Visani, M., Hoang, V.D., 2018. Deep cnn and data augmentation for skin lesion classification, in: Asian Conference on Intelligent Information and Database Systems, Springer. pp. 573–582.

Qin, Z., Liu, Z., Zhu, P., Xue, Y., 2020. A gan-based image synthesis method for skin lesion classification.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Computer Methods and Programs in Biomedicine , 105568.

Quang, N.H., et al., 2017. Automatic skin lesion analysis towards melanoma detection, in: 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), IEEE. pp. 106–111.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv:1711.05225 .

Rawat, W., Wang, Z., 2017. Deep convolutional neural networks for image classification: A comprehensive review. Neural computation 29, 2352–2449.

Rebouças Filho, P.P., Peixoto, S.A., da Nóbrega, R.V.M., Hemanth, D.J., Medeiros, A.G., Sangaiah, A.K., de Albuquerque, V.H.C., 2018. Automatic histologically-closer classification of skin lesions. Computerized Medical Imaging and Graphics 68, 40–54.

Rehman, M.u., Khan, S.H., Rizvi, S.D., Abbas, Z., Zafar, A., 2018. Classification of skin lesion by interference of segmentation and convolotion neural network, in: 2018 2nd International Conference on Engineering Innovation (ICEI), IEEE. pp. 81–85.

Ries, L.A., Harkins, D., Krapcho, M., Mariotto, A., Miller, B., Feuer, E.J., Clegg, L.X., Eisner, M., Horner, M.J., Howlader, N., et al., 2006. Seer cancer statistics review, 1975-2003 .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. arXiv:1609.04747 .

Serte, S., Demirel, H., 2019. Gabor wavelet-based deep learning for skin lesion classification. Computers in biology and medicine 113, 103423.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35, 1285–1298.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of Big Data 6, 60.

Siegel, R.L., Miller, K.D., Jemal, A., 2020. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians 70, 7–30.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 .

Smith, L., MacNeil, S., 2011. State of the art in non-invasive imaging of cutaneous melanoma. Skin Res. Technol. 17, 257–269.

Smith, L.N., 2017. Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

on Applications of Computer Vision (WACV), IEEE. pp. 464–472.

Song, L., Lin, J.P., Wang, Z.J., Wang, H., 2020. An end-to-end multi-task deep learning framework for skin lesion analysis. IEEE Journal of Biomedical and Health Informatics .

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

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

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

Talo, M., Baloglu, U.B., Yıldırım, Ö., Acharya, U.R., 2019. Application of deep transfer learning for automated brain abnormality classification using mr images. Cognitive Systems Research 54, 176–188.

Tang, P., Liang, Q., Yan, X., Xiang, S., Zhang, D., 2020a. Gp-cnn-dtel: Global-part cnn model with data-transformed ensemble learning for skin lesion classification. IEEE Journal of Biomedical and Health Informatics .

Tang, Y., Fang, Z., Yuan, S., Xing, Y., Zhou, J.T., Yang, F., et al., 2020b. imscgnet: Iterative multi-scale context-guided segmentation of skin lesion in dermoscopic images. IEEE Access 8, 39700–39712.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 180161.

Valle, E., Fornaciali, M., Menegola, A., Tavares, J., Bittencourt, F.V., Li, L.T., Avila, S., 2020. Data, depth, and design: Learning reliable models for skin lesion analysis. Neurocomputing 383, 303–313.

Xie, F., Yang, J., Liu, J., Jiang, Z., Zheng, Y., Wang, Y., 2020. Skin lesion segmentation using high-resolution convolutional neural network. Computer Methods and Programs in Biomedicine 186, 105241.

Yadav, S.S., Jadhav, S.M., 2019. Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data 6, 113.

Yang, X., Zeng, Z., Yeo, S.Y., Tan, C., Tey, H.L., Su, Y., 2017. A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv:1703.01025 .

Yilmaz, E., Trocan, M., 2020. Benign and malignant skin lesion classification comparison for three deep-

medRxiv preprint doi: https://doi.org/10.1101/2021.02.02.21251038; this version posted February 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

learning architectures, in: Asian Conference on Intelligent Information and Database Systems, Springer. pp. 514–524.

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A., 2016. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging 36, 994–1004.

Yu, Z., Jiang, F., Zhou, F., He, X., Ni, D., Chen, S., Wang, T., Lei, B., 2020. Convolutional descriptors aggregation via cross-net for skin lesion recognition. Applied Soft Computing , 106281.

Yu, Z., Jiang, X., Zhou, F., Qin, J., Ni, D., Chen, S., Lei, B., Wang, T., 2018. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. IEEE Transactions on Biomedical Engineering 66, 1006–1016.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method. arXiv:1212.5701 .

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE transactions on medical imaging 38, 2092–2103.

Zhang, N., Cai, Y.X., Wang, Y.Y., Tian, Y.T., Wang, X.L., Badami, B., 2020. Skin cancer diagnosis based on optimized convolutional neural network. Artificial Intelligence in Medicine 102, 101756.

Zhu, L., Feng, S., Zhu, W., Chen, X., 2020. Asnet: An adaptive scale network for skin lesion segmentation in dermoscopy images, in: Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, International Society for Optics and Photonics. p. 113170W.

Zucker, S.W., Terzopoulos, D., 1980. Finding structure in co-occurrence matrices for texture analysis. Computer graphics and image processing 12, 286–308.