Computerized Medical Imaging and Graphics 88 (2021) 101843

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images

Imran Iqbal [a], Muhammad Younus [b], Khuram Walayat [c], Mohib Ullah Kakar [d], Jinwen Ma [a,*]

[a] Department of Information and Computational Sciences, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, People's Republic of China
[b] State Key Laboratory of Membrane Biology and Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine and Peking-Tsinghua Center for Life Sciences and PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, People's Republic of China
[c] Faculty of Engineering Technology, Department of Thermal and Fluid Engineering, University of Twente, Enschede, 7500 AE, Netherlands
[d] Beijing Key Laboratory for Separation and Analysis in Biomedicine and Pharmaceuticals, Beijing Institute of Technology, Beijing, 100081, People's Republic of China

ARTICLE INFO

Keywords:
Artificial intelligence
Computer vision
Convolutional neural network
Deep learning
Dermoscopy
Image processing
Melanomas
Nevi
Pattern recognition
Skin cancer screening
Skin lesion classification

ABSTRACT

As an analytic tool in medicine, deep learning has gained great attention and opened new ways for disease diagnosis. Recent studies validate the effectiveness of deep learning algorithms for binary classification of skin lesions (i.e., melanomas and nevi classes) with dermoscopic images. Nonetheless, those binary classification methods cannot be applied to the general clinical situation of skin cancer screening in which multi-class classification must be taken into account. The main objective of this research is to develop, implement, and calibrate an advanced deep learning model in the context of automated multi-class classification of skin lesions. The proposed Deep Convolutional Neural Network (DCNN) model is carefully designed with several layers, and multiple filter sizes, but fewer filters and parameters to improve efficacy and performance. Dermoscopic images are acquired from the International Skin Imaging Collaboration databases (ISIC-17, ISIC-18, and ISIC-19) for experiments. The experimental results of the proposed DCNN approach are presented in terms of precision, sensitivity, specificity, and other metrics. Specifically, it attains 94 % precision, 93 % sensitivity, and 91 % specificity in ISIC-17. It is demonstrated by the experimental results that this proposed DCNN approach outperforms state-of-the-art algorithms, exhibiting 0.964 area under the receiver operating characteristics (AUROC) in ISIC-17 for the classification of skin lesions and can be used to assist dermatologists in classifying skin lesions. As a result, this proposed approach provides a novel and feasible way for automating and expediting the skin lesion classification task as well as saving effort, time, and human life.

1. Introduction

Cancer refers to a disease caused by the uncontrolled growth of abnormal cells in the body and often has the potential to replicate, divide, spread through the lymph and blood, and destroy normal body tissues ("National Cancer Institute. (2015). What is Cancer?," 2015). Its mortality rate is the second highest after cardiovascular disease in the world. According to the International Agency for Research on Cancer (IARC), more than 9 million patients died and over 18 million new cases of cancer were reported worldwide in 2018 (Cancer - World Health Organization [WWW Document], 2018). Environmental factors such as air pollution, family history, and poor lifestyle choices such as alcohol and smoking can damage deoxyribonucleic acid (DNA) that may lead to cancer. It is clear that there is still a long way to effectively controlling the mortality of cancer. However, with the help of fast development of image processing and artificial intelligence (AI) algorithms for diagnosis and prognosis of the diseases, the chances of surviving many forms of cancer are increasing considerably in recent years.

There are six main classes of cancer: 1) Carcinoma is a cancer that originates in the skin, pancreas, lungs, breasts, and other organs and glands; 2) Sarcoma arises in the bone, cartilage, muscle, fat, blood vessels, or other connective tissues of the body; 3) Leukemia begins in the blood-forming tissue, such as the bone marrow, and causes large number of abnormal blood cells to be produced; 4) Lymphoma is a cancer that develops in the cells of the immune system; 5) Central nervous system cancer starts in the tissues of the spinal cord and brain; 6)

* Corresponding author.
E-mail addresses: imraniqbalrajput@pku.edu.cn (I. Iqbal), younusmuhamamd@pku.edu.cn (M. Younus), k.walayat@utwente.nl (K. Walayat), mohibullah44@yahoo.com, mohib.kakar@bit.edu.cn (M.U. Kakar), jwma@math.pku.edu.cn (J. Ma).

https://doi.org/10.1016/j.compmedimag.2020.101843
Received 5 September 2020; Received in revised form 13 November 2020; Accepted 11 December 2020
Available online 24 December 2020

Melanoma is a type of skin cancer that begins in the cells that make the pigment in skin and that can spread to other organs.

Skin cancer is one of the most widespread and fatal cancer types globally (Karimkhani et al., 2017). It is a key health concern with over 10,000 newly reported cases every month around the world (Harangi, 2018). It generally develops due to exposure to ultraviolet (UV) rays from the sun, which harms the DNA of skin cells. Some artificial sources of light, in particular tanning beds and sunlamps, increase the risk of developing this disease. Genetic defects are also a main source of this type of cancer (The Skin Cancer Foundation, 2018).

Skin lesion can be categorized into several classes, including Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic Keratosis (AK), Benign keratosis lesion (BKL), Dermatofibroma (DF), Vascular lesion (VASC), and Squamous cell carcinoma (SCC). BCC and SCC are most often found in the areas exposed to the sun, such as the head, neck, and arms. Most of the skin cancer classes are very common and also remediable. MEL is more likely to grow and spread than the other types of skin cancer. MEL represents less than 5% of all skin cancer forms, however, it is held responsible for over 70 % of all the fatalities caused by skin cancer (Kanimozhi and Murthi, 2016). If MEL is classified correctly in the early stages, the probability of mortality of patients could be decreased (Jerant et al., 2000). Manual recognition of MEL needs experienced dermatologists to overcome the problems of high degree of inter-class similarities and intra-class differences of skin lesions. Consequently, if the MEL classification has been performed automatically, it will improve accuracy and efficiency of the early detection of this type of skin cancer (Hosny et al., 2019).

Currently, the examination of skin cancer is performed visually by clinical experts. In fact, clinical screening is the preliminary analysis, which is followed by biopsy, histopathological testing, and dermoscopic assessment (Esteva et al., 2017). In fact, attributive classification of skin lesions plays a critical role in the early and accurate diagnosis of skin cancer. However, it requires specific proficiency that might not be available in general clinical settings. Dermoscopy is the examination of skin via skin surface microscopy, essentially for evaluating pigmented skin lesions. This skin imaging modality has been developed to assist dermatologists and improve diagnostic accuracy in contrast to unaided visual inspection (Kittler et al., 2002; Vestergaard et al., 2008). The classification of skin lesions is particularly based on color features, dermal features, contour features, geometric features, and texture features of lesions. The visual classification of skin lesions is difficult and may lead to wrong recognition of lesions considering the high degree of visual similarities among different lesion classes (Codella et al., 2015). For that reason, classification of skin lesions through deep convolutional neural network (DCNN) is an effective and alternative solution of the visual examination. From International Skin Imaging Collaboration-2019 (ISIC-19) (Malvehy et al., 2019) and recent studies (Chaturvedi et al., 2019; Esteva et al., 2017; Gessert et al., 2020; Harangi, 2018; Hekler et al., 2019; Hosny et al., 2019; Liu et al., 2020; Mahbod et al., 2019; Rebouças Filho et al., 2018; Winkler et al., 2020), it has been found that the classification of skin lesions with deep learning is still a challenging task due to the following reasons: 1) The classes are highly imbalanced (e.g., the NV class has ~54 times more examples than the DF class); 2) There is a high degree of inter-class similarities as well as intra-class differences; 3) Dermoscopic images contain various artifacts including hair, gel bubbles, ruler markers, ink markers, color illumination, patches, ebony frames, and blood vessels which make the recognition task very challenging.

Recent studies show significant performance of binary classification of skin lesions with deep learning models (Esteva et al., 2017). Nonetheless, these models cannot apply to the general multi-class classification of skin lesions with similar classification performance. The main aim of this research is to develop, implement, and calibrate an advanced deep learning model in the context of multi-class classification of skin lesions with minimal pre-processing operations. This specialized DCNN model is designed to accurately classify dermoscopic skin lesion images into multiple classes. This approach is good to expedite the automated multi-class classification process of skin lesions. It owns the competency of deep learning that exceeds dermatologists in terms of accuracy and throughput. The experimental results demonstrate that our proposed DCNN-based approach has a potential to assist dermatologists for classifying dermoscopic skin lesion images.

2. Related works

Skin cancer is a common human malignancy (Rogers et al., 2015; Society, 2016; Stern, 2010) which is mostly diagnosed visually by clinical experts, starting with a primary clinical screening and followed by dermoscopic assessment, a biopsy, and histopathological testing. Various algorithms of classification of skin lesions utilize conventional artificial intelligence methods, which normally begin with a phase of handcrafted feature extraction, followed by a separate training phase of the classifier. Earlier approaches are based on low-level hand-crafted features for classification of MEL and non-melanoma skin lesions (Barata et al., 2019). Handcrafted features generally suffer from poor generalization capability for dermoscopic images because of obscure understanding of biological mechanisms and involving weak human intuition. Thus, the low-level handcrafted features are not suitable for distinguishing complex skin lesion images. The authors (Celebi et al., 2007) proposed their method to select the hand-crafted features but those features have huge visual resemblance issues, high degree of intra-class differences, and artifacts appearance in dermoscopic images gave poor performance. Esteva et al. (Esteva et al., 2017) employed Inception-v3 architecture to trained on a dataset having more than 100,000 clinical images annotated by experienced dermatologists. They showed their method based on DCNN was capable to surpass 21 board-certified dermatologists in terms of classifying the skin cancer through dermoscopy and photographic imaging if the training set was adequately large. For getting a high area under the receiver operating characteristics (AUROC) using dermoscopic images, Balazs Harangi (Harangi, 2018) proposed an ensemble technique based on neural network for skin lesion classification. He fused the outputs of the classification layers of four network architectures and showed that his approach achieved better result than the individual model. His proposed model attained 0.89 average AUROC for three skin lesion classes. Mahbod et al. (Mahbod et al., 2019) proposed algorithm that utilized ensembles learning and pre-trained network to classify skin lesions and attained competitive results to an experienced dermatologist. Their algorithm achieved specificity 78 % for MEL and 86 % for BKL classes. Gessert et al. (Gessert et al., 2020) proposed a patch-based attention method for skin lesion classification task of high-resolution images with three pre-trained networks. They used approaches such as balanced batch sampling, class-specific loss weighting, and oversampling to tackle the skewed class distributions issue. Moreover, they also proposed a diagnosis-guided loss technique which considers the algorithm used for ground-truth annotation. In recent work, Liu et al. (Liu et al., 2020) proposed a two phase approach consists of mid-level feature. In the first phase, they detected the region of interest of dermoscopic images, while the second phase of their approach based on pre-trained models to extract features for region of interest images. Their mid-level feature based algorithm attained 0.87 AUROC for MEL and 0.97 AUROC for BKL classes.

3. Methodology

3.1. Pre-processing, partitioning, and augmentation of datasets

Dermoscopic images are acquired from ISIC-17, ISIC-18, and ISIC-19 databases for experiments. In fact, ISIC-19 is comprised of HAM10000 (Tschandl et al., 2018), BCN (Combalia et al., 2019), and MSK (Codella et al., 2018) datasets. There are a total of 25,331 labeled images which are publicly available for classification task with eight labels: MEL, NV,

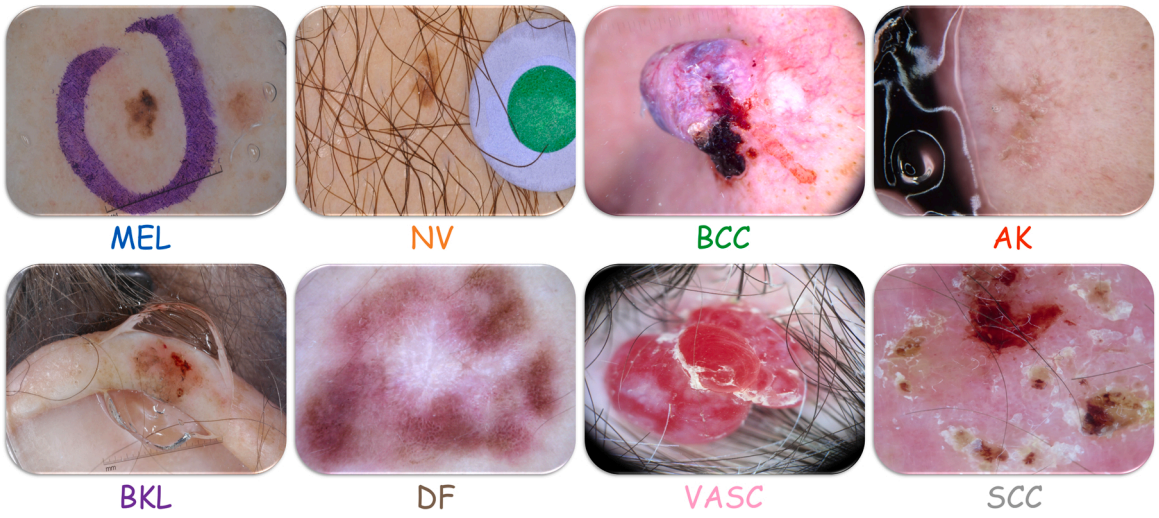

Fig. 1. Typical samples of skin lesion of each class in ISIC-19 with various artifacts such as hair artifact (NV, BKL, DF, VASC, SCC), ink marker artifact (MEL), ruler marker artifact (MEL, BCC, BKL), gel bubble artifact (MEL, AK, BKL), and patch artifact (NV).

Table 1
ISIC-19 description, partitioning, and augmentation. "W", "H", and "C" signify the width, height, and color channels of image, respectively.

Classes	Training set (~70 %)	Augmented training set	Development set (~10 %)	Test set (~20 %)	Total (100 %)	Average resolution of original image (W×H×C)
MEL	3166	9498	452	904	4522	917 × 844 × 3
NV	9013	9013	1287	2575	12,875	801 × 677 × 3
BCC	2326	9304	332	665	3323	958 × 935 × 3
AK	607	9105	87	173	867	960 × 938 × 3
BKL	1837	9185	262	525	2624	846 × 736 × 3
DF	167	9018	24	48	239	820 × 748 × 3
VASC	177	9027	25	51	253	786 × 702 × 3
SCC	440	9240	63	125	628	891 × 844 × 3
Total or (average)	17,733	73,390	2532	5066	25,331	(855 × 761 × 3)

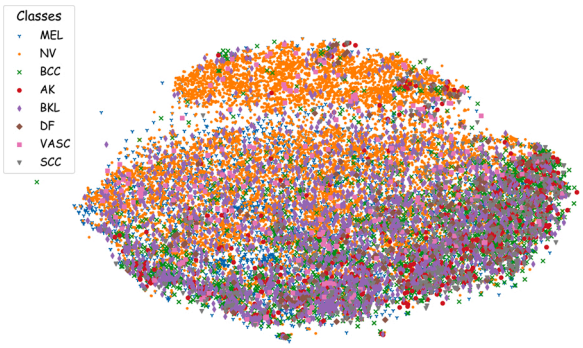

Fig. 2. Visualization of the samples of eight skin lesion classes in ISIC-19 by the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm.

BCC, AK, BKL, DF, VASC, and SCC. Moreover, there are 101 different resolution images ranging from 576 × 768–1024 × 1024 with 3 color channels. The number of samples of the NV class has ~54 times more than DF class samples. For illustration, Fig. 1 shows some typical samples of ISIC-19 with various artifacts.

In the pre-processing step, cropping is applied to the dermoscopic images in such a way that each image is transformed into a square image and the center of the lesion appears in the center of the corresponding image. During the pre-processing, the aspect ratio of each image is preserved. Every image is rescaled to a resolution of 64 × 64 pixels using inter nearest interpolation to retain its information and reduce the computational cost of processing. OpenCV library is employed for the pre-processing of the dermoscopic images. There is no need to remove the artifacts in the pre-processing step from the dermoscopic images such as hair, gel bubbles, ruler markers, ink markers, patches, dark borders, and others artifacts because the proposed DCNN model is very smart and intelligent to cope these types of artifacts with ease. ISIC-19 is partitioned into three parts, where the first part roughly contains 70 % of the data from each class to form the training set, the second part contains about 10 % images to form the development set for tuning the hyper-parameters of the proposed model (see Table 1 for more details), while the remaining part contains about 20 % of the data from each class to form the test set. Fig. 2 shows the result of visualization for the eight skin lesion classes of ISIC-19 by the T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm.

In order to solve the problem of skewed classes, overfitting, and training image scarcity, more augmentation operations are implemented to the underrepresented classes, while less or no augmentation operations to the overrepresented classes to balance the sample size in each class of the training set. Thus, the training set is extended virtually with the actual classes being balanced. For instance, the BCC and SCC classes have 2326 and 440 distinct images, respectively, but the sample sizes of the two corresponding augmented classes are similar, i.e., 9304 and 9240, respectively (see Table 1). As for the precise data augmentation, three common techniques are adopted: rotation, translation, and flipping. Specifically, an image is rotated by -30 to 30 degrees. For translation, an image is shifted 12.5 % to the left, the right, up, and down. For flipping, an image is flipped horizontally and vertically. It should be noted that these augmentation operations are only implemented on the training set, while the development and test sets only contain the

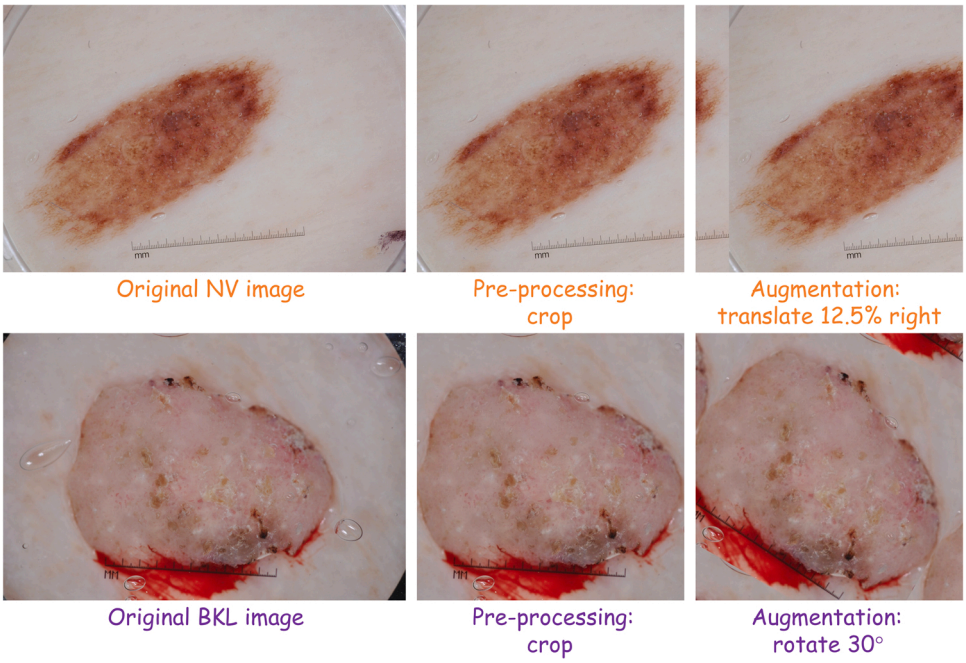

Fig. 3. Typical original, pre-processed, and augmented samples of Melanocytic nevus (NV) and Benign keratosis lesion (BKL) classes in ISIC-19.

original images. Fig. 3 shows some typical NV and BKL samples using the pre-processing and augmentation operations with the border wrap function of OpenCV library.

In ISIC-18, there are a total of 10,015 labeled images which are publicly available for classification task with seven labels: MEL, NV, BCC, AK, BKL, DF, and VASC. The number of samples of the NV class has ~58 times more than DF class samples. In ISIC-17, there are a total of 2750 labeled images (training set: 2000, development set: 150, and test set: 600) for classification tasks with three labels: MEL, NV, and BKL. The number of samples of the NV class has ~5 times more than BKL class samples. There are two binary classification tasks of skin lesions in ISIC-17. To distinguish MEL with NV and BKL skin lesions in the first task. And distinguishing BKL with MEL and NV classes in the second task. Similar pre-processing and augmentation operations, mentioned above, are also applied to ISIC-17 and ISIC-18, however, for partitioning, similar settings are utilized to directly compare the performance of proposed DCNN model with the previous methods such as Lina Liu et al. settings for ISIC-17 and Nils Gessert et al. settings for ISIC-18.

3.2. Proposed DCNN model

Being inspired by advanced DCNN models (He et al., 2016; Huang et al., 2017; Iqbal et al., 2020a, 2020b; Szegedy et al., 2017), a specialized DCNN model is proposed for the skin lesion classification that has been a challenging task even for experienced dermatologists. To tackle this problem, the proposed DCNN network, being named as Classification of Skin Lesions Network (CSLNet), utilizes four key kernel units as shown in Fig. 4. The first to fourth units are based on Block C with 3, 6, 9, and 3 repetitions, respectively, from top to bottom in the top left subfigure. Moreover, they are linked by Block D. In fact, Block C is a composition of Block A, B, and their concatenation operation is shown in the bottom subfigure, while Block A, B, and D are shown in the top right corner. The number of filters in Block A and B are 128 and 32, where their filter sizes are 1×1 and 3×3, respectively. The number of filters in Block D is equal to half the number of existing channels. In the first unit, there are 9 convolutional layers, while the second unit consists of 18 convolutional layers, and the third kernel unit comprises 27 convolutional layers. These units may detect the degree of symmetry and color such as black-superficial epidermis, brown–epidermis, grey-papillary dermis, and blue-reticular dermis and responsible to detect the complex patterns such as reticular, globular, homogenous, parallel, cobblestone, lacuna, and starburst patterns, and may extract the complex lesion features such as atypical pigment network, streaks, dots and globules, pigment blotch, and milia-like cysts. In the last unit, there are further 9 convolutional layers to learn the features that are quite precise to describe the classes of skin lesions. The number of filters utilized by Amirreza Mahbod et al., Balazs Harangi, and Lina Liu et al. approaches are ~58.3 K, ~84.7 K, and ~29.1 K, respectively, while the number of parameters used by these approaches are ~256.7 M, ~267.5 M, and ~45.6 M, respectively. In contrast, the number of filters and number of parameters used by the proposed DCNN model are ~4.6 K and ~4.8 M, respectively.

Normally, the original images are rescaled to a lower resolution for training the network since the computational power and memory are generally limited. This rescaling procedure implies that the fine-grained image contains enough information in a medical context such as skin lesion classification task. While designing the proposed model, we also carefully remain aware of the issues of this classification problem such as skewed lesion class distributions, high degree of inter-class similarities and intra-class variabilities, and presence of artifacts in dermoscopic images, which actually make the recognition task very challenging. Thus, we utilize 68 convolutional layers in the proposed DCNN model. Before each convolutional layer, the batch normalization (Ioffe and Szegedy, 2015) and LeakyReLU (Maas et al., 2013) are applied. This DCNN model is trained by the backpropagation algorithm that is actually based on the gradient descent rule of the loss with respect of the weights in the network. LeCun normal initializers (Klambauer et al., 2017) are used to initialize the biases and weights. LeakyReLU and softmax are employed as the activation functions for the convolutional layers and output layer, respectively. L2 norm is used as the kernel regularizer in a dense layer to prevent overfitting. The stochastic gradient descent is performed with Adamax optimizer (Kingma and Ba, 2015) to update the weights of the proposed network with a mini batch size of 512 for 60 epochs. The learning rate is set by 0.0007. Furthermore, the categorical cross entropy is utilized as the loss function. Experiments are performed using Keras (Chollet, 2017) with TensorFlow

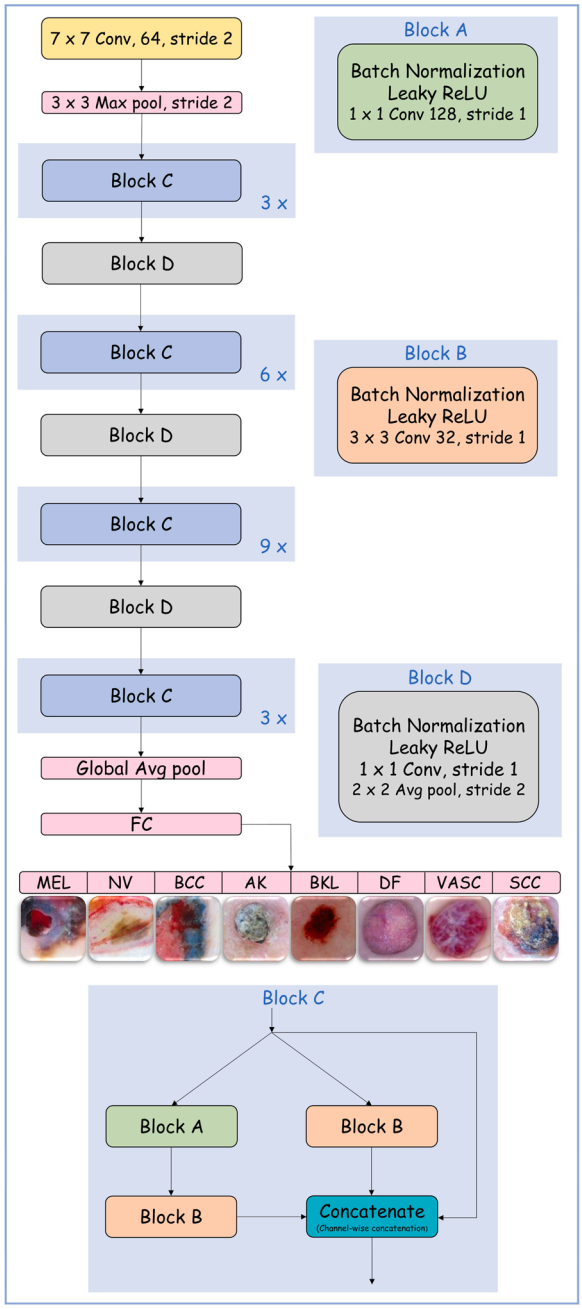

Fig. 4. The layout of the proposed network, CSLNet, where "FC" signifies the fully connected layer.

(Abadi et al., 2016) backend on GPU. The hyperparameters of the proposed model are tuned on the development set of ISIC-19. Specifically, the hyperparameters are selected according to the lowest loss of the DCNN model evaluated on the development set. Finally, the obtained model is assessed on the test set.

4. Experiments and results

In this section, several experiments of the proposed DCNN model, CSLNet, were performed for the classification of skin lesions with the dermoscopic images. It was tested on ISIC-17, ISIC-18, and ISIC-19, and compared with the state-of-the-art methods. Precision, sensitivity, specificity, accuracy, F1 score, Jaccard similarity coefficient (JSC), geometric mean (G-mean), Matthews correlation coefficient (MCC), Cohen's kappa score (CKS), AUROC, precision-recall curve (PR-AUC), and evaluation time were considered as the metrics for the classification evaluation. Several earlier methods needed extensive pre-processing and extraction of domain-precise visual features before classification. In the contrary, CSLNet did not need any hand-crafted features and was trained end-to-end directly with the dermoscopic images to classify skin lesions. Specifically, it was implemented with the TensorFlow and Keras framework on a NVIDIA GeForce GTX 1080 card with 8GB GDDR5X memory. The training process took roughly 73 min for ISIC-19.

We firstly trained CSLNet on the augmented training set and tuned the hyperparameters on the development set. There were 2532 skin lesion images in the development set in ISIC-19. This sample size was similar to the standard of the ImageNet computer vision challenge (Russakovsky et al., 2015), which has 50–100 images per object category in its development set. In fact, hyperparameters were tuned to improve the performance of CSLNet and Table 2 shows the near optimal values of the hyperparameters, which were selected according to the best performance of the proposed model on the development set. That is, the model of CSLNet with the lowest loss on the development set was chosen for evaluation on the test set.

The experimental results of CSLNet are shown in Figs. 5–8. Six main evaluation indices are used to evaluate the performance of the proposed and comparative approaches such as precision, sensitivity, specificity, accuracy, F1 score, and AUROC. Fig. 5a-d show the classification accuracy and loss curves with the number of epochs during the training on the training and test sets of ISIC-17, ISIC-18, and ISIC-19. It can be seen that the training process converged within 60 epochs. Notably, CSLNet surpassed the existing deep learning-based skin lesion classification approaches in terms of the precision, sensitivity, specificity, accuracy, F1 score, and AUROC, which are clearly shown in Table 3, where the bold values denote the best results. Only the specificity for ISIC-17 of the Rebouças Filhoa et al. method is slightly better (~1.6 %) than the specificity of the proposed DCNN model, whereas the specificity of the proposed DCNN model is similar to the specificity of the method proposed by Nils Gessert et al. for ISIC-18.

The confusion matrices were calculated (shown in Fig. 6a-d), which provide a good insight on how often images of each individual class (MEL, NV, BCC, AK, BKL, DF, VASC, and SCC) are correctly classified or misclassified by the proposed model on the test set. Element (i, j) of confusion matrix represents the empirical probability of predicting class j given that the ground truth is class i. After carefully examining the confusion matrices (Fig. 6c and d), we can find that the NV class is very difficult to distinguish from the remaining classes. The main reason for this may be that the NV class has a variety of forms. In contrast, we also find that the average true positive rate (TPR) of the BKL class is relatively high so that the BKL images can be easily recognized. As mentioned before, skin lesion classes specially NV has huge intraclass variation, and there is a high degree of visual similarity between this and the other lesions, which may affect the classification performance.

We further plotted the precision-recall curves of classes on the test set as well as their micro-averaging precision-recall curve (shown in Fig. 7a-d), where a large area under the precision-recall curve signifies

Table 2
Hyperparameter configurations of the proposed DCNN model.

DCNN	Learning algorithm	Learning rate	Mini-batch size	Epochs	Regularizer	Activation function	Data augmentation
Proposed network (CSLNet)	Adamax	7e-4	512	60	L2	LeakyReLU	flip, translation, rotation

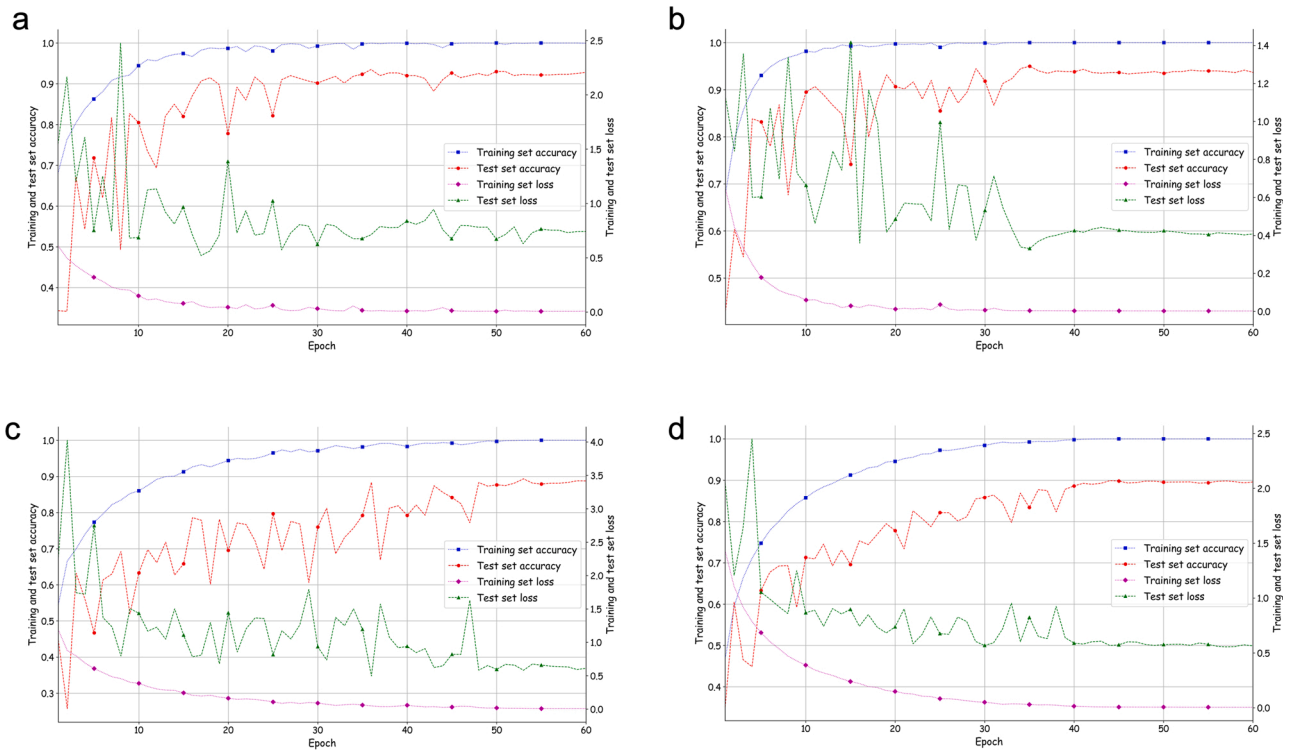

Fig. 5. Classification accuracy and loss curves of CSLNet with the number of epochs during the training on the training and test sets. (a) MEL vs NV and BKL skin lesion classes in ISIC-17. (b) MEL and NV vs BKL skin lesion classes in ISIC-17. (c) Seven skin lesion classes in ISIC-18. (d) Eight skin lesion classes in ISIC-19.

the high precision as well as the high recall. It can be observed from Fig. 7c and 7d that the NV class has the highest PR-AUC, whereas the MEL and VASC classes have the lowest one in ISIC-18 and ISIC-19, respectively. Finally, we plotted the receiver operating characteristic curves of classes on the test set as well as their macro and micro-averaging receiver operating characteristic curves (shown in Fig. 8a-d). The receiver operating characteristic curve is also valuable because it shows the tradeoff between the TPR and false positive rate (FPR). From Fig. 8c and 8d, we can further find that the AK and DF classes have the highest AUROC in ISIC-18 and ISIC-19, respectively, whereas the NV class has the lowest one.

Precision, sensitivity, specificity, accuracy, F1 score, AUROC, PR-AUC, MCC, CKS, JSC, and G-mean of proposed DCNN model in the first binary classification task, MEL vs NV and BKL, of ISIC-17 are 93.56 %, 92.83 %, 91.39 %, 92.83 %, 93.05 %, 0.952, 0.944, 0.791, 0.786, 0.874, and 0.920, respectively, while these metrics for second binary classification task, MEL and NV vs BKL, of ISIC-17 are 94.38 %, 93.67 %, 89.90 %, 93.67 %, 93.89 %, 0.977, 0.974, 0.775, 0.770, 0.889, and 0.920, respectively. The overall performance of proposed DCNN model in terms of these metrics in ISIC-17 are 93.97 %, 93.25 %, 90.64 %, 93.25 %, 93.47 %, 0.964, 0.959,0.783, 0.778, 0.882, and 0.920, respectively. Precisions, sensitivities, accuracies, and F1 scores of the proposed model in ISIC-17, ISIC-18, and ISIC-19 are quite similar, while the specificity in ISIC-17 is lower than the specificities in ISIC-18 and ISIC-19. Training and evaluation time of the proposed DCNN model is much faster than the methods proposed by Amirreza Mahbod et al. and Lina Liu et al. We also calculated MCC, CKS, JSC, and G-mean metrics of the proposed DCNN model for future comparison. After carefully examining Table 3, we can clearly see that the proposed DCNN model not only perform better for binary classification tasks but also work well for multi-class classification problems.

5. Conclusions and final remarks

Skin cancer is a leading health problem all over the world and skin lesion classification has a major role in the early and accurate diagnosis of skin cancer. In order to improve the classification performance, we have established a specialized DCNN model, CSLNet, for automated multi-class classification of skin lesions with dermoscopic images. Based on the skin lesion images from ISIC, data-driven deep learning algorithms can be utilized to solve this challenging problem. The classification of skin lesions is an intricate problem owning to its intrinsic inter-class similarities and intra-class variabilities. By making careful pre-processing and augmentation of ISIC-17, ISIC-18, and ISIC-19 images, we design a specialized DCNN model for the classification of skin lesions. The TPR of eight classes (Fig. 6d) indicates that it is reliable in recognizing the images in the BKL class as well as the other classes. Our proposed approach has achieved 90 % accuracy, 91 % precision, 98 % specificity, and 90 % sensitivity in ISIC-19. These results strongly support the statement that AI algorithms are useful to medical specialists during the interpretation of medical imaging. Our proposed approach obtains good results even with the small size of the images i.e., $64 \times 64 \times 3$. Its evaluation time is ~0.2 milliseconds (ms) per image. By attaining the domain experts-level classification performance, it can also be a balanced classifier where the TPR is similar to the positive predictive value (PPV). It achieves better classification results than the previous state-of-the-art methods without using transfer leaning. From Table 3, it can be clearly seen that the proposed DCNN model achieves better results in all the metrics than the comparative methods except that the specificities of the approaches proposed by Rebouças Filhoa et al. for ISIC-17 and Nils Gessert et al. for ISIC-18.

Developing an automated classification system of skin lesions can greatly reduce the workload of dermatologists and also decrease the subjectivity and inaccuracy of the classification task induced by human error. Due to incorrect or late diagnosis, a few cases of wrong treatment have been reported. Since effect of treatment takes time to appear,

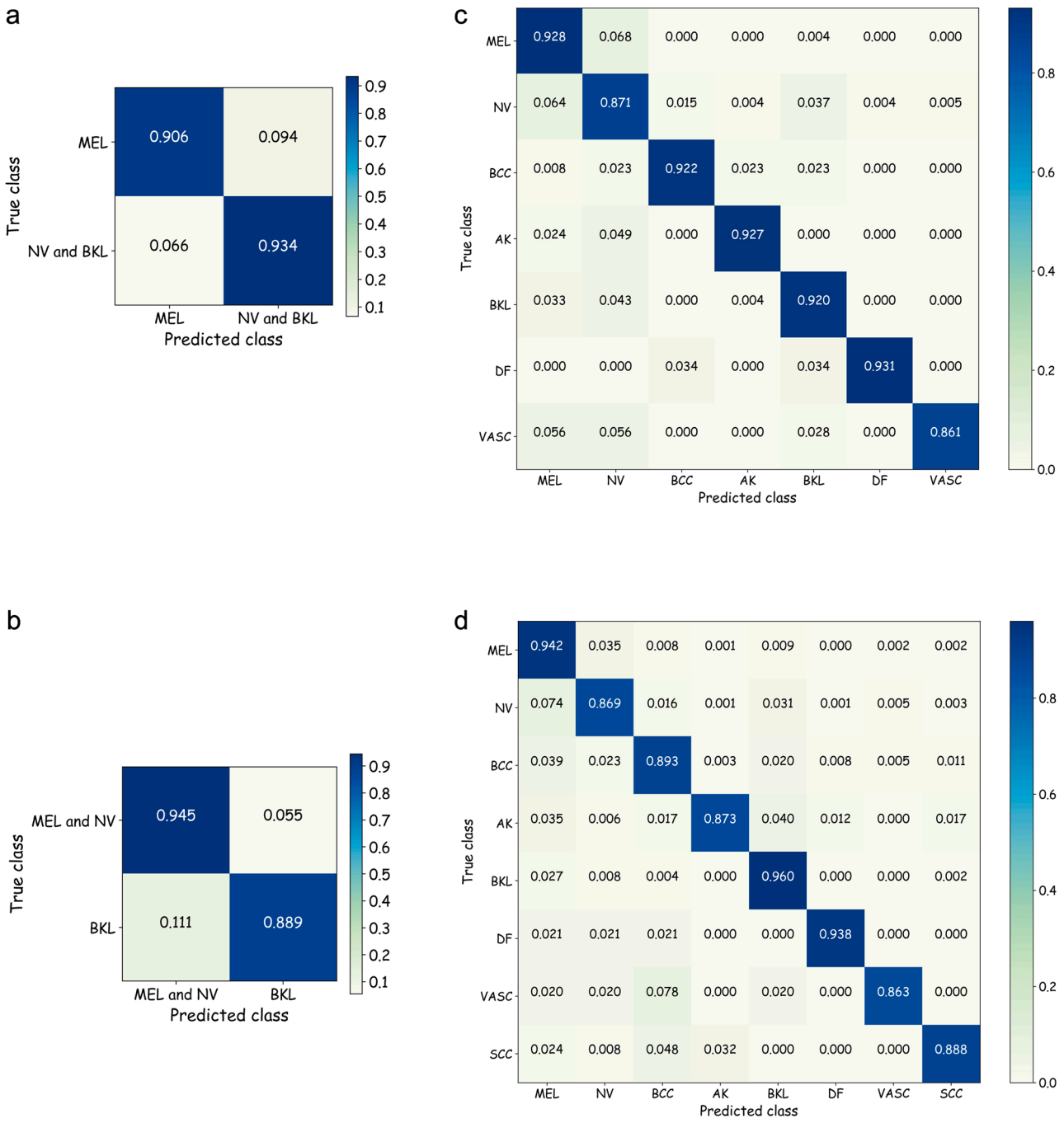

Fig. 6. CSLNet confusion matrices on the test sets. (a) MEL vs NV and BKL skin lesion classes in ISIC-17. (b) MEL and NV vs BKL skin lesion classes in ISIC-17. (c) Seven skin lesion classes in ISIC-18. (d) Eight skin lesion classes in ISIC-19.

sometimes misdiagnosis leads to an increased need for surgical treatment and hospitalization duration (Martin C. McHenry et al., 2002). The attainment of skin cancer examination is considerably reliant on the diagnostic proficiency of dermatologists conducting the skin inspection. Nonetheless, dermatologists who have more than ten years' experience hardly surpass 80 % recall, while dermatologists who have three to five years' experience attain only 62 % recall in the skin cancer screening (Morton and Mackie, 1998; Vestergaard et al., 2008). This proposed system can become crucial and more valuable when experienced dermatologists are not readily available and for inexperienced clinicians in the underdeveloped countries. It can even be applied to assign a class label to new skin lesion and this DCNN model is good to expedite the automated classification procedure of skin lesions. While these results are encouraging and provide strong support to that the deep learning approach is able to play a key role in assisting doctors and healthcare systems, but further validation and improvement are still required with more clinical information of subjects such as age, gender, race, and family history to evaluate the deep models in clinical practice.

Author contributions

I.I. and J.M. devised the project, the main conceptual ideas and proof

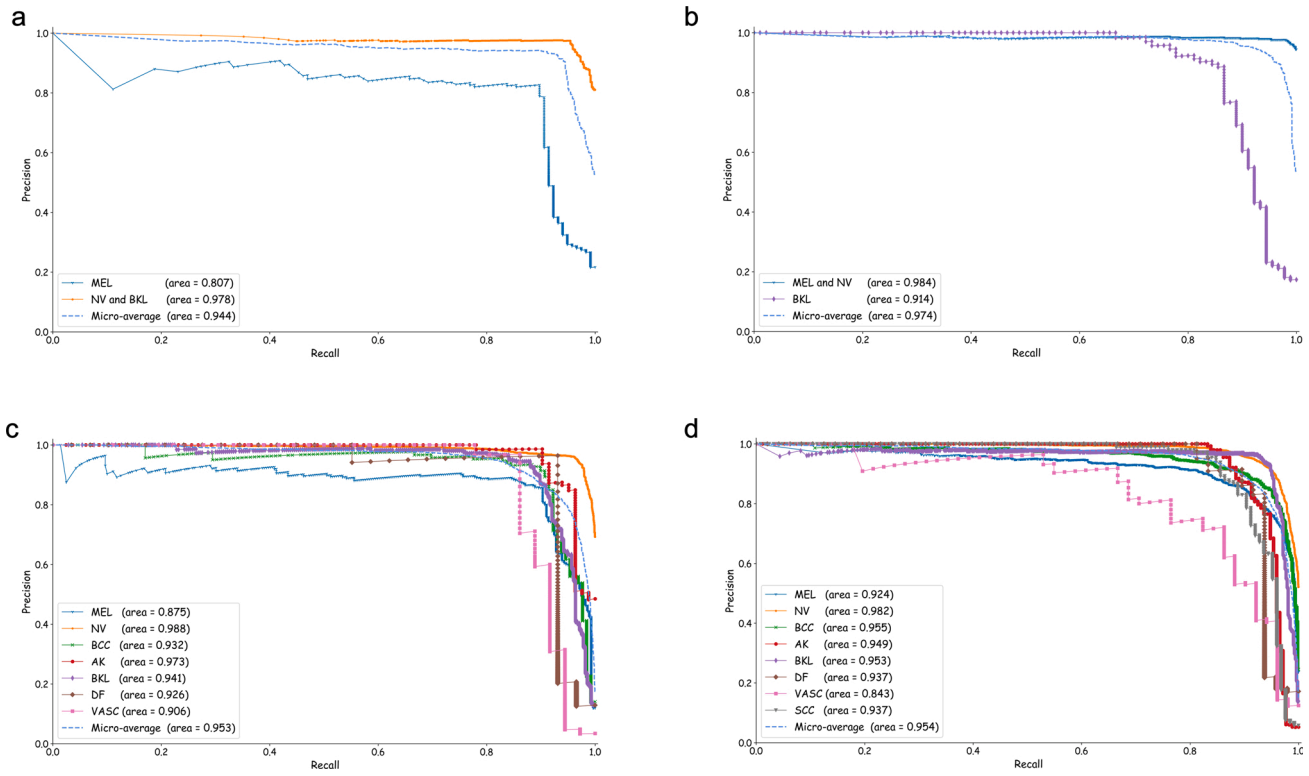

Fig. 7. CSLNet precision-recall curves on the test sets as well as micro-averaging precision-recall curve. (a) MEL vs NV and BKL skin lesion classes in ISIC-17. (b) MEL and NV vs BKL skin lesion classes in ISIC-17. (c) Seven skin lesion classes in ISIC-18. (d) Eight skin lesion classes in ISIC-19.

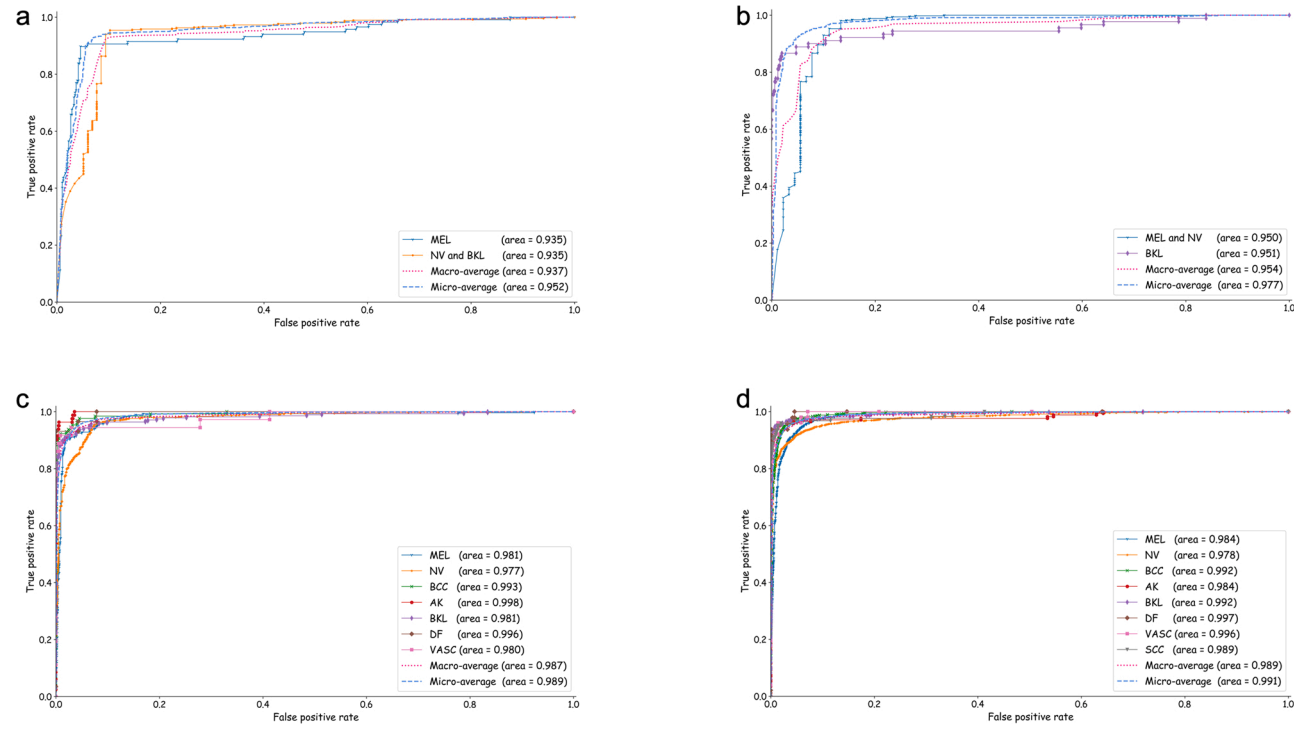

Fig. 8. CSLNet receiver operating characteristic curves on the test sets as well as macro and micro-averaging receiver operating characteristic curves. (a) MEL vs NV and BKL skin lesion classes in ISIC-17. (b) MEL and NV vs BKL skin lesion classes in ISIC-17. (c) Seven skin lesion classes in ISIC-18. (d) Eight skin lesion classes in ISIC-19.

Table 3
The performance of our proposed and comparative methods in ISIC-17, ISIC-18, and ISIC-19, where all the metrics except MCC, CKS, AUROC, PR-AUC, JSC, and G-mean are in percent. Bold font shows the best results. The symbol "-" stands for unreported results. In three columns (8th, 9th and 10th), six metrics are presented to shorten the width of the table.

Methods/Authors	Dataset	Precision	Sensitivity	Specificity	Accuracy	F1 score	AUROC (PR-AUC)	MCC (CKS)	JSC (G-mean)	Training time (minutes) (Evaluation time (milliseconds per image))
Matsunaga et al.	ISIC-17	–	–	–	81.6	–	0.911 (-)	–	–	–
Gonzalez-Diaz		–	–	–	84.9	–	0.910 (-)	–	–	–
Menegola et al.		–	–	–	88.3	–	0.908 (-)	–	–	–
Rebouças Filhoa et al.		91.29	89.93	**92.15**	89.93	90.0	0.890 (-)	–	–	–
Balzas Harangi		–	55.6	78.5	86.6	–	0.891 (-)	–	–	–
Amirreza Mahbod et al.		–	87.26	82.18	87.7	–	0.914 (-)	–	–	415 (-)
Lina Liu et al.		–	–	–	–	–	0.921 (-)	–	–	96 (390)
Proposed network (CSLNet)		**93.97**	**93.25**	90.64	**93.25**	**93.47**	**0.964** (0.959)	0.783 (0.778)	0.882 (0.920)	**~24 × 2 (~0.2)**
Nils Gessert et al.	ISIC-18	–	75.7	**96.0**	–	82.6	–	–	–	–
Saket S. Chaturvedi et al.		89	83	–	83.15	83	–	–	–	–
Proposed network (CSLNet)		**90.45**	**88.75**	95.72	**88.75**	**89.11**	0.989 (0.953)	0.807 (0.800)	0.806 (0.922)	~25 (~0.2)
Proposed network (CSLNet)	ISIC-19	90.66	89.58	97.57	89.58	89.75	0.991 (0.954)	0.854 (0.851)	0.815 0.933	~73 (~0.2)

outline; I.I. designed and performed the experiments with support from K.W; I.I. drafted the manuscript, designed the figures and tables with support from M.Y and M.U.K.; I.I. and J.M. contributed to the interpretation of the results; J.M. supervised the project; All the authors have read and approved the final submitted manuscript.

Funding

This work was supported by the National Key Research and Development Program of China under grant 2018AAA0100205

Declaration of Competing Interest

The authors declare no conflict of interest.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Barata, C., Celebi, M.E., Marques, J.S., 2019. A survey of feature extraction in dermoscopy image analysis of skin Cancer. IEEE J. Biomed. Heal. Informatics 23, 1096–1109. https://doi.org/10.1109/JBHI.2018.2845939.

Cancer - World Health Organization [WWW Document], 2018. Int. Agency Res. Cancer. URL https://www.who.int/cancer/PRGlobocanFinal.pdf (accessed 2.17.20).

Celebi, M.E., Kingravi, H.A., Uddin, B., Iyatomi, H., Aslandogan, Y.A., Stoecker, W.V., Moss, R.H., 2007. A methodological approach to the classification of dermoscopy images. Comput. Med. Imaging Graph. 31, 362–373. https://doi.org/10.1016/j.compmedimag.2007.01.003.

Chaturvedi, S.S., Gupta, K., Prasad, P.S., 2019. Skin Lesion Analyser: An Efficient Seven-Way Multi-class Skin Cancer Classification Using MobileNet. https://doi.org/10.1007/978-981-15-3383-9_15.

Chollet, F., 2017. Keras (2015). URL http//keras. io..

Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R., 2015. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. Machine Learning in Medical Imaging. Springer, Munich, Germany, pp. 118–126. https://doi.org/10.1007/978-3-319-24888-2_15.

Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A., 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). International Symposium on Biomedical Imaging, Washington, USA. https://doi.org/10.1109/ISBI.2018.8363547.

Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Halpern, A.C., Puig, S., Malvehy, J., 2019. BCN20000: Dermoscopic Lesions in the Wild.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118. https://doi.org/10.1038/nature21056.

Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A., 2020. Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. IEEE Trans. Biomed. Eng. 67, 495–503. https://doi.org/10.1109/TBME.2019.2915839.

Harangi, B., 2018. Skin lesion classification with ensembles of deep convolutional neural networks. J. Biomed. Inform. 86, 25–32. https://doi.org/10.1016/j.jbi.2018.08.006.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. Computer Vision and Pattern Recognition, Las Vegas, USA, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hekler, A., Utikal, J.S., Enk, A.H., Solass, W., Schmitt, M., Klode, J., Schadendorf, D., Sondermann, W., Franklin, C., Bestvater, F., Flaig, M.J., Krahl, D., von Kalle, C., Fröhling, S., Brinker, T.J., 2019. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur. J. Cancer 118, 91–96. https://doi.org/10.1016/j.ejca.2019.06.012.

Hosny, K.M., Kassem, M.A., Foaud, M.M., 2019. Classification of skin lesions using transfer learning and augmentation with Alex-net. PLoS One 14. https://doi.org/10.1371/journal.pone.0217293.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Conference on Computer Vision and Pattern Recognition. Honolulu, USA, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: 32nd International Conference on Machine Learning, ICML 2015. Lille, France, pp. 448–456.

Iqbal, I., Mustafa, G., Ma, J., 2020a. Deep learning-based morphological classification of human sperm heads. Diagnostics 10, 325. https://doi.org/10.3390/diagnostics10050325.

Iqbal, I., Shahzad, G., Rafiq, N., Mustafa, G., Ma, J., 2020b. Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning. IET Image Process. https://doi.org/10.1049/iet-ipr.2019.1646.

Jerant, A.F., Johnson, J.T., Sheridan, C.D., Caffrey, T.J., 2000. Early detection and treatment of skin cancer. Am. Fam. Physician 62, 357–386.

Kanimozhi, T., Murthi, A., 2016. Computer aided Melanoma skin cancer detection using Artificial Neural Network classifier. Singaporean J. Sci. Res. J. Sel. Areas Microelectron. 8, 35–42.

Karimkhani, C., Green, A.C., Nijsten, T., Weinstock, M.A., Dellavalle, R.P., Naghavi, M., Fitzmaurice, C., 2017. The global burden of melanoma: results from the Global Burden of Disease Study 2015. Br. J. Dermatol. 177, 134–140. https://doi.org/10.1111/bjd.15510.

Kingma, D.P., Ba, J.L., 2015. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations. San Diego, USA.

Kittler, H., Pehamberger, H., Wolff, K., Binder, M., 2002. Diagnostic accuracy of dermoscopy. Lancet Oncol. 3, 159–165. https://doi.org/10.1016/S1470-2045(02)00679-4.

Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S., 2017. Self-normalizing neural networks. In: Advances in Neural Information Processing Systems. CA, USA.

Liu, L., Mou, L., Zhu, X.X., Mandal, M., 2020. Automatic skin lesion classification based on mid-level feature learning. Comput. Med. Imaging Graph. 84 https://doi.org/10.1016/j.compmedimag.2020.101765.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning. Atlanta, Georgia.

Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Pitiot, A., Wang, C., 2019. Fusing fine-tuned deep features for skin lesion classification. Comput. Med. Imaging Graph. 71, 19–29. https://doi.org/10.1016/j.compmedimag.2018.10.007.

Malvehy, J., Halpern, A., Codella, N.C.F., Celebi, M.E., Combalia, M., Gutman, D., Helba, B., Kittler, H., Rotemberg, V., Tschandl, P., 2019. Skin Lesion Analysis Towards Melanoma Detection (ISIC 2019) [WWW Document]. URL https://challenge2019.isic-archive.com (accessed 2.17.20).

McHenry, Martin C., Easley, K.A., Locker, G.A., 2002. Vertebral osteomyelitis: long-term outcome for 253 patients from 7 cleveland-area hospitals. Clin. Infect. Dis. 34, 1342–1350. https://doi.org/10.1086/340102.

Morton, C.A., Mackie, R.M., 1998. Clinical accuracy of the diagnosis of cutaneous malignant melanoma. Br. J. Dermatol. 138, 283–287. https://doi.org/10.1046/j.1365-2133.1998.02075.x.

National Cancer Institute, 2015. What Is Cancer? [WWW Document], 2015. URL https://www.cancer.gov/about-cancer/understanding/what-is-cancer (accessed 2.17.20).

Rebouças Filho, P.P., Peixoto, S.A., Medeiros da Nóbrega, R.V., Hemanth, D.J., Medeiros, A.G., Sangaiah, A.K., de Albuquerque, V.H.C., 2018. Automatic histologically-closer classification of skin lesions. Comput. Med. Imaging Graph. 68, 40–54. https://doi.org/10.1016/j.compmedimag.2018.05.004.

Rogers, H.W., Weinstock, M.A., Feldman, S.R., Coldiron, B.M., 2015. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. JAMA Dermatol. 151, 1081–1086. https://doi.org/10.1001/jamadermatol.2015.1187.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Society, A.C., 2016. Cancer Facts & Figures 2016 [WWW Document]. Atlanta, Am. Cancer Soc. URL. http://www.cancer.org/acs/groups/content/@research/%0Adocuments/document/acspc-047079.pdf%0A.

Stern, R.S., 2010. Prevalence of a history of skin cancer in 2007: results of an incidence-based model. Arch. Dermatol. 146, 279–282. https://doi.org/10.1001/archdermatol.2010.4.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI, pp. 4278–4284 https://doi.org/arXiv:1602.07261.

The Skin Cancer Foundation, 2018. Skin Cancer Information. [WWW Document]. URL https://www.skincancer.org/skin-cancer-information (accessed 2.17.20).

Tschandl, P., Rosendahl, C., Kittler, H., 2018. Data descriptor: the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5. https://doi.org/10.1038/sdata.2018.161.

Vestergaard, M.E., Macaskill, P., Holt, P.E., Menzies, S.W., 2008. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br. J. Dermatol. 159, 669–676. https://doi.org/10.1111/j.1365-2133.2008.08713.x.

Winkler, J.K., Sies, K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Abassi, M.S., Fuchs, T., Rosenberger, A., Haenssle, H.A., 2020. Melanoma recognition by a deep learning convolutional neural network—performance in different melanoma subtypes and localisations. Eur. J. Cancer 127, 21–29. https://doi.org/10.1016/j.ejca.2019.11.020.