Tissue and Cell 58 (2019) 76–83

Contents lists available at ScienceDirect

Tissue and Cell

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

A comparative study of deep learning architectures on melanoma detection

Sara Hosseinzadeh Kassani[a,*], Peyman Hosseinzadeh Kassani[b]

[a] Department of Computer Science, University of Saskatchewan, Saskatchewan, Canada
[b] Department of Biomedical Engineering, Tulane University, New Orleans, Louisiana, USA

ARTICLE INFO

Keywords:
Cancer classification
Computational diagnosis
Convolutional neural networks
Deep learning
Melanoma detection

ABSTRACT

Melanoma is the most aggressive type of skin cancer, which significantly reduces the life expectancy. Early detection of melanoma can reduce the morbidity and mortality associated with skin cancer. Dermoscopic images acquired by dermoscopic instruments are used in computational analysis for skin cancer detection. However, some image quality limitations such as noises, shadows, artefacts exist that could compromise the robustness of the skin image analysis. Hence, developing an automatic intelligent system for skin cancer diagnosis with accurate detection rate is crucial. In this paper, we evaluate the performance of several state-of-the-art convolutional neural networks in dermoscopic images of skin lesions. Our experiment is conducted on a graphics processing unit (GPU) to speed up the training and deployment process. To enhance the quality of images, we employ different pre-processing steps. We also apply data augmentation methodology such as horizontal and vertical flipping techniques to address the class skewness problem. Both pre-processing and data augmentation could help to improve the final accuracy.

1. Introduction

Computational analysis for inferring the skin lesion data is an area of increased research interest because of its importance in cancer diagnosis and treatment planning. Melanocytes are melanin-producing cells whose main purpose is to give color to skin, hair and eyes (Almasni et al., 2018; Flores and Scharcanski, 2016). Skin cancer, the most common type of cancer (Xu et al., 2018), can be classified into two categories: melanoma and non-melanoma (Okur and Turkan, 2018). The high morbidity and considerable healthcare cost associated with the malignant type of lesion (Fig. 1a) has motivated researchers to develop more accurate and flexible algorithms for early melanoma detection (Pathan et al., 2018) since, abnormal melanocyte cells divide out of control and spread to other tissues of the body and become highly metastasis and increase the number of deaths. According to the data provided by American Cancer Society (American Cancer Society, 2018), the rates of melanoma have been increasing during the last 30 years. In 2018, about 91,270 new melanoma cases will be diagnosed and about 9320 people are expected to die of melanoma in the United States. However, melanoma is highly curable if diagnosed in its early stage. Fig. 1 illustrates some examples of skin lesions.

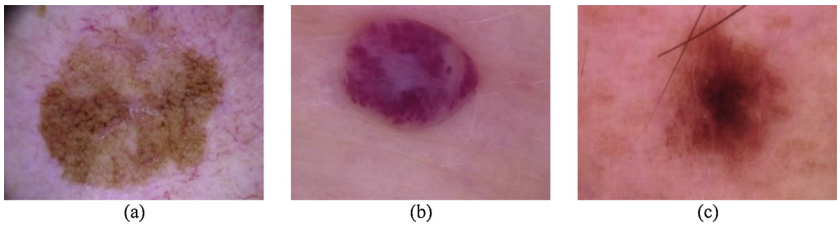
Employing effective non-invasive diagnostic techniques aid dermatologists to collect data, providing insights to melanoma structure and shape. Macroscopic images, better known as non-dermoscopic or clinical images - since such images captured using conventional digital cameras are inexpensive and non-invasive - used in computational analysis for patients of different ages and ethnicities (Oliveira et al., 2016a). Dermoscopic images acquired by dermoscope, a special non-invasive imaging instrument, generally have better invariance to illumination and contrast (Bi et al., 2018). However, different image acquisition conditions and parameters such as variable distances, non-ideal illumination, poor resolution, and also image quality limitations such as noises, shadows, artefacts, hairs and reflections decrease the image quality that could compromise the robustness of the skin lesion analysis (Jafari et al., 2016). Therefore, a computer aided diagnosis system that is able to analyze dermoscopic-based images may be take into consideration to aid dermatologist's in melanoma diagnosis (Oliveira et al., 2016a).

For skin lesion image analysis in this paper, the image pre-processing steps are fundamental in skin lesion detection. Different image pre-processing step including color space correction, contrast enhancement method and noise reduction can effectively improve the prediction of skin cancer (Okur and Turkan, 2018). The objective of this study is to apply the state-of-the-art deep learning architectures on dermoscopic images to classify skin lesions.

The rest of the paper is organized as follows: A review of previous related works, motivations and contributions of this study are presented in Section 2. A detailed description of data augmentation and

* Corresponding author.
E-mail address: sara.kassani@usask.ca (S. Hosseinzadeh Kassani).

https://doi.org/10.1016/j.tice.2019.04.009
Received 3 March 2019; Accepted 19 April 2019
Available online 22 April 2019
0040-8166/



(a) (b) (c)

Fig. 1. (a) Malignant skin lesion (b) Vascular lesions (c) Melanocytic Nevus lesions.

preprocessing steps, candidate Deep Convolutional Neural Networks (DCNN) and evaluation metrics is proposed in Section 3. The experimental data, results and discussion of comparative study are reported in Section 4, and finally Section 5 draws the conclusion.

2. Related works

During the last few years, advances in deep convolutional neural networks have shown promising results in object recognition tasks and became a demanding research domain for classification in medical image processing (Naylor et al., 2017). Most of the studies in melanoma detection involve only machine learning algorithms, however, there are a few thoroughly skin lesions classification studies using deep learning algorithms.

Instead of training a CNN from scratch with randomly initialized parameters, Kawahara et al. (2016) utilizeed a pretrained CNN to classify skin images on the entire dataset. This pre-training highly reduceed the training time of CNN and achieves 85.8% accuracy over 5-classes.

Liao's work (Liao, 2016) provided an universal skin disease classification by using transfer learning strategy and fine-tuning ImageNet pretrained weights on the Dermnet dataset to improve the training performance. They evaluate the model performance on both the Dermnet dataset and skin images from New York State Department of Health skin disease dataset. Their method achieved 73.1% Top-1accuracy and 91.0% Top-5 accuracy on the Dermnet dataset. and 69.5% Top-5 accuracies on the OLE dataset.

Ercal et al. (1994) developed a method to detect skin tumors using color features. The accuracy reported by this approach is about 82%. The highest success rate is achieved for “intradermal nevus” at 100% and the lowest success rate is for “melanoma” images at 77%.

Menegola et al. (2017) proposed knowledge transfer method to enhance performance of deep learning for melanoma screening. In their study, a pre-trained model trained on the Kaggle Challenge for Diabetic Retinopathy Detection dataset. They expected that the deeper models and transfer learning from a related dataset improve the performance and leads to better results. However, their work suggest that the experimental design is sensitive to the type of skin lesions (benign or malignant).

Lopez et al. (2017) emolyed a deep-learning based approach for early detection of melanoma. Their solution is based on a modified VGGNet architecture and transfer learning technique to solve the skin lesion classification task. The proposed method achieveed a sensitivity value of 78.66% on the ISIC Archive dataset.

In a study done by Ayan, (2018), the performance of a CNN architecture is compared between a non-augmented dataset and augmented dataset for classification of skin lesions. They proposed that the data augmentation methods could be useful for building powerful classifiers with insufficient data. Results showed that the network using the augmented dataset has achieved better accuracy rate than non-augmented data.

Erçal et al. (1999), introduced a skin cancer classification using diagnostic-tree based hierarchical neural networks and fuzzy logic based on morphological features. There were four classes in their study. These classes are: malignant melanoma, atypical mole, basal cell carcinoma or actinic keratosis, and intradermal nevus or seborrheic keratosis.

Fatima et al. (2012) proposed a methodology based on a Multi-Parameter Extraction and Classification System (MPECS) of twenty-one features using six phases and then analyzed them based on statistical methods for early detection of skin cancer melanoma. Some of the features are lesion borders, color, symmetry, area, perimeter and the eccentricity. They used Sobel edge detection for detecting lesion edges. The symmetry feature is computed by horizontal and vertical axis of lesions. The color spreading factor of the lesions is computed based on the similarity of neighborhood pixels. Their study demonstrated singular statistical analysis from extracted features is not adequately sufficient to accurately classify the skin lesions. Therefore, advanced classification methods such as support vector machines and decision tree algorithms could be considerate for classification.

We tabulate some of the recent studies on Melanoma detection at Table 1.

2.1. Motivations and contributions

With above review in place, we found that accurate detection of skin cancer is crucial for providing necessary treatment for patients. Several studies show that developing accurate methods by employing deep learning algorithm plays an emerging role in the diagnosis of various diseases such as cancer (Hu et al., 2018), pneumonia (Kermany et al., 2018), and Alzheimer (Amoroso et al., 2018). Interestingly, study at Dorj et al. (2018), shows high accuracies obtained for skin cancer classification using AlexNet (Krizhevsky et al., 2012). Studies at Pereira et al. (2016), Vasconcelos and Vasconcelos, (2017) and Wahab et al. (2017) show that data augmentation and pre-processing may further help to get a more accurate and unbiased detector. According to the observations of the study at Lei et al. (2018), the highest accuracy, 93.46% is gained by a modified ResNet50 model with a deeper model,

Table 1
List of recent studies on Melanoma detection.

Dataset	Methods	(Pre-processing, augmentation)	Accuracy	Year of publication
PH2	Satheesha et al. (2017)	(+, -)	95.75	2017
	Bi et al. (2016)	(-, -)	92	2016
	Waheed et al. (2017)	(-, -)	96	2017
	Barata et al. (2014)	(+, -)	84.32	2014
ISIC 2016	Codella et al., 2019	(-, -)	91.6	2017
	Matsunaga et al. (2017)	(-, -)	83.09	2017
	Yu et al. (2017)	(+, +)	85	2017
	Lopez et al. (2017)	(+, +)	81.33	2017

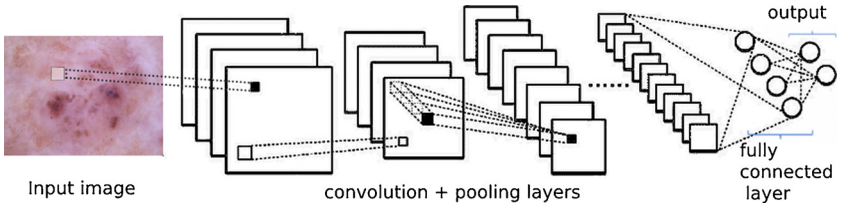


Fig. 2. The architecture of a typical convolution neural network.

i.e. using more hidden layers in the structure of DCNN. So, all these studies motivated us to examine the performance of five state of the arts DCNNs, namely AlexNet (Krizhevsky et al., 2012), VGGNet16 (Simonyan and Zisserman, 2014), VGGNet19 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2016), and Xception (Chollet, 2017) with the aim of verifying the effectiveness of various pre-processing and data augmentation techniques.

In the following, we list contributions of this study:

1 We added two fully connected layers to combines the feature maps of various intermediate layers of DCNNs. For both layers, dropout method is employed which avoids the overfitting problem. Moreover, the strategy of transferring features learned from ImageNet dataset (which is different form medical datasets) is used. To minimize the training loss, *L2* regularization (ridge regression) technique is also employed for classification task. These changes could bring higher classification performance to our learners.
2 To reduce heterogeneity of data, and hence, improvement of the classification performance, the *pre-processing* steps such as contrast enhancement, color space transformation, and a specific type of illumination correction proposed by Cavalcanti et al. at (Cavalcanti et al., 2010) are used. These preprocessing steps also help to reduce the sensitivity of the DCNNs to contrast and intensity bias. Therefore, the DCNNs are able to learn high-level structure information more efficiently and ensure to gain a better prediction.
3 The dataset provided for this study is severely skewed. To balance the dataset, an *oversampling* method for minority classes examples is proposed. This method consists of different data augmentation techniques such as horizontal and vertical flips, random contrast and random brightness that aid to avoid the negative effect of class imbalances.

3. Material and methodology

In this section we briefly explain the pre-processing and data augmentation methods, DCNNs, and evaluation metrics of learners. All these networks are compared with each other and also with and without preprocessing and augmentation steps.

3.1. Pre-processing

The key point of pre-processing is to reduce the effect of noise and unbalanced patterns in the dataset that tend to decrease the ability of DCNNs in learning important features. Common pre-processing steps are contrast enhancement, color space transformation, and illumination correction (Oliveira et al., 2016b). Also, non-skin noises such as hair, and rulers should be eliminated from skin images. There are also noisy representations such as uneven illumination patterns, reflections, poor background, different lightening and shadows that adversely affect image quality.

3.2. Data augmentation

Oversampling and undersampling are common methods to fix imbalanced datasets. Oversampling is utilized to generate either exact copies, or modified copies of the original samples that belongs to the minority class to adjust the class distribution. There are different data augmentation techniques by the number of random images transformation such as horizontal and vertical flips, random contrast and random brightness. These techniques help to avoid the negative effect of strong class imbalances and hence increase the performance of DCNNs.

3.3. Deep convolutional neural networks

DCNNs have achieved the state-of-the-art performance in many applications such as image and speech recognition (Fayek et al., 2017), automatic video classification (Shao et al., 2014), object detection (Pathak et al., 2018) and natural language processing (Sun et al., 2017). A Convolution Neural Network (CNN), known as ConvNet, is a specific type of feed-forward neural network with a stack of convolutional layers, each followed by pooling layers in order to extract features from the input data and produce a set of high-level feature maps at each level of convolution. The feature maps information is summarized using pooling layers in order to reduce the number of parameters and further is followed by a fully connected layer to produce the final classification (Sainath et al., 2013; Krizhevsky et al., 2012). Fig. 2 illustrates a graphical representation of a typical convolution neural network.

3.4. Xception architecture

The Xception deep neural network which stands for extreme inception, was made by François Chollet Chollet (2017). Xception architecture has depth-wise separable convolutions. Xception has 36 convolutional layers to extract important features and is inspired by Inception Chollet (2017) wherein the Inception modules are replaced with depth-wise separable convolutions consisting of a depth-wise convolution.

3.5. AlexNet architecture

AlexNet (Krizhevsky et al., 2012) architecture won the first prize in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 competition. AlexNet is the first breakthrough in the architecture of CNNs applied to large datasets. The basic architecture of AlexNet consists of five convolutional layers, two normalization layers, three max-pooling layers, three fully-connected layers, and a linear layer with Softmax activation function in the output to ensure each neuron predicts class probabilities of a particular image.

3.6. VGGNet architecture

VGGNet was introduced in 2014 by Karen Simonyan and Andrew Zisserman (Simonyan and Zisserman, 2014) from Visual Geometry Group (VGG) of the University of Oxford. It achieves one of the top performances in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014. VGGNet utilizes smaller filters of 3×3, compared to AlexNet 11×11 filter, in order to provide better features extraction from images. The study also verifies that using much smaller filters in order to increase the depth of network instead of its width plays a critical role for gaining higher performance. There are two versions of this architecture: VGG16 and VGG19 with different depths and layers. VGG19 is deeper than VGG16. However, the number of parameters for VGG19 is larger and is hence more expensive to train the

Table 2
Total number of training and test images before and after augmentation.

	BCC	BKL	MEL	NV	VASC	DF	AKIEC
Original Training Data	247	815	801	4975	109	85	253
Augmented Training Data	4773	3275	3308	1111	3488	4152	3856
Original Test Data	139	284	312	1730	33	30	74
Augmented Test Data	1692	1479	1664	1111	1484	1632	1552

network compared to VGG16.

3.7. ResNet architecture

Deep residual neural network (ResNet) architecture is proposed by He et al. (2016) and won ILSVRC & COCO 2015 competitions. Researchers are able to train deeper and more effective neural networks using ResNet with better recognition accuracies. ResNets with various depths such as ResNet50 and ResNet100 use the bottleneck features to improve efficiency in compare with its predecessor CNN models (Lei et al., 2018).

3.8. Evaluation metrics

To compare the performance of the DCNNs architectures for skin lesions prediction task, we employ various standard evaluation metrics such as specificity, sensitivity, accuracy, and F-measure. According to Eqs. (1)–(5), true positives (TP) is the numbers of instances that correctly predicted; false negatives (FN) is the numbers of instances that incorrectly predicted. True negatives (TN) is the numbers of negative instances correctly predicted, while false positives (FP) is the numbers of negative instances incorrectly predicted (Yang et al., 2017). Sensitivity or recall is the measure of skin lesion labels that correctly classified. Sensitivity is critical specially in the medical field and is given by:

$$\text{Sensitivity or True Positive Rate(TPR)} = \frac{TP}{(TP + FN)} \tag{1}$$

Specificity is the measure of non-skin lesion labels that successfully classified and is expressed as:

$$\text{Specificity or True Negative Rate(TNR)} = \frac{TN}{(TN + FP)} \tag{2}$$

Precision or positive predictive value measures what percentage of correctly classified labels is truly positive and is given as:

$$\text{Positive Predictive Value(PPV)} = \frac{TP}{(TP + FP)} \tag{3}$$

Accuracy (ACC) is used to show the number of correctly classified skin lesions divided by the total number of skin lesions and is defined as:

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{4}$$

F1 score, also known as F-measure is defined as the weighted average of precision and recall that combines both the precision and recall together. F-measure is expressed as,

$$F - measure = 2 \times \frac{(precision \times recall)}{(precision + recall)} \tag{5}$$

4. Result and discussion

4.1. Dataset and experimental setup

The dataset used for this research is skin lesion analysis towards melanoma detection provided by International Skin Imaging Collaboration (ISIC, 2018) which is publicly available at (ISIC, 2018). The goal of this challenge is to develop computer analysis systems that assist dermatologists for skin lesions detection from dermoscopic images. The challenge is divided into three separate tasks i) lesion boundary segmentation, ii) lesion attribute detection and iii) lesion diagnosis. The training dataset consisted of 10,015 dermoscopic images. The size of each images is 600 × 450 pixels and a range of diseases vary from benign to malignant. Possible disease classes are: 1) Melanoma, 2) Melanocytic nevus, 3) Basal cell carcinoma, 4) Actinic keratosis / Bowen's disease (intraepithelial carcinoma), 5) Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis), 6) Dermatofibroma, 7) Vascular lesion.

Some experimental setups are as follows: To evaluate the performance of different architectures, 70% of images of each class are randomly chosen for training and the remaining 30% for test. There are no common images between the training and test sets for each class of images. The training parameters for all networks are, learning rate $\eta = e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = e^{-8}$, dropout rate and batch size are set to 0.5 and 32, respectively. To prevent overfitting, we set regularization parameter to be 0.0001. The momentum rate and the weight decay parameters are set to 0.9 and e^{-5}, respectively. We also used ImageNet pretrained weights since the use of pretrained layers leads to more robust performance. The class distribution of skin lesions is highly imbalanced, i.e. 68%, 11% and 10% for the cases NV, BKL and MEL respectively before dividing the data into test and train set. We tabulated the class distributions in Table 2. As demonstrated in Table 2, BCC, AKIEC, VASC and DF classes have few examples.

To avoid the negative effect of strong class imbalances and increase the performance of DCNN, we performed different data augmentation techniques by different image transformations such as horizontal and vertical flips, random contrast and random brightness. To improve the contrast and resolution of skin lesion images, we used shades-of-grey method, proposed by Barata et al (Simonyan and Zisserman, 2014). For correcting illumination variation, a method proposed by Cavalcanti et al. (2010) is used. The illumination variation correction is achieved by a quadratic function and mainly summarized as:

- Convert the original input image from RGB color space to the HSV color space.
- Apply a quadratic function computed from the local illumination intensity to the V channel in HSV color space.
- Convert back the image from the HSV color space to the RGB color space.

We scaled all images to the size of 224 × 224 pixels for AlexNet model and 225 × 300 for Xception, VGGNet16, VGGNet19 and ResNet50. Then, for image normalization, we rescaled the intensity values of the pixels to be at the range [0, 1]. To remove bias from the features, we standardize data with removal of the mean, making dataset with a zero mean and a standard deviation equal to one.

The network architectures are implemented in Python using the Keras package with Tensorflow backend on a Intel(R) Core(TM) i7-8700 K 3.7 GHz processors and NVIDIA GTX 1080 TI with 11 GB graphical processing unit (GPU) and 32GB RAM.

Some examples of data augmentation steps are illustrated in Fig. 3. After data augmentation the size of dataset increases by 4 times with 27,827 images for training set and 11,233 images for test set.

4.2. Overall performance

The results obtained from candidate DCNN architectures are presented in Table 3 for Melanoma classification. The results show that ResNet50 architecture with data augmentation and pre-processing steps achieves the best accuracy rate in the majority of the validation processes and it has the highest average F-score of 92.74%, and accuracy of 92.08%.

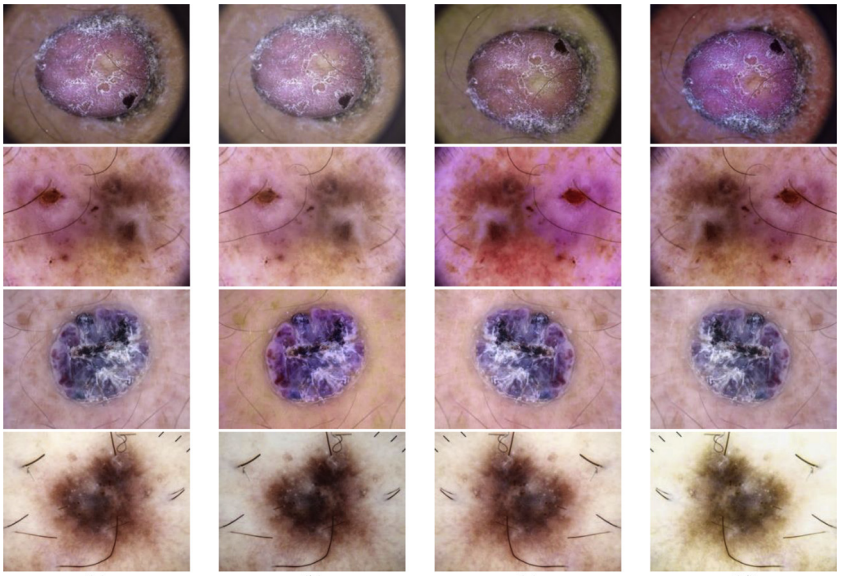


Fig. 3. pre-processed and augmented data. (a) Original dermoscopy image, (b) Image after application of illumination correction, (c) Image after application of contrast enhancement space and horizontal flip, (d) Image after transformation into HSV color space and horizontal flip.

Table 3
Evaluation results of the DCNN classification models (The sizes of images used to train AlexNet is 224 × 224 × 3 and for Xception, VGGNet16, VGGNet19 and ResNet50 are 225 × 300 × 3. An asterisk beside the model name indicates that data were preprocessed and augmented.

Methods	Precision	Recall	F-Score	ACC
AlexNet	0.7717	0.7873	0.7726	0.7853
AlexNet*	0.8421	0.8125	0.8231	0.8045
ResNet50	0.8652	0.8663	0.8537	0.8637
ResNet50*	**0.9373**	**0.9253**	**0.9274**	**0.9208**
VGGNet16	0.8442	0.8447	0.8433	0.8436
VGGNet16*	0.907	0.9032	0.9061	0.8836
VGGNet19	0.8468	0.8457	0.8436	0.8461
VGGNet19*	0.8855	0.8882	0.8838	0.8870
Xception	0.4472	0.6633	0.5346	0.6629
Xception*	0.9019	0.9057	0.9041	0.9030

the best values under the same conditions are shown boldface.

The asterisk (*) indicates that the model was trained using the data augmentations and preprocessing steps as described in 4.1. ResNet50*, i.e. after augmentation and preprocessing, has 92% accuracy while AlexNet*, and VGG19* have accuracy 80% and 88% respectively. This means the gap in accuracy is 12% and 4% in favor of ResNet50* compared to the AlexNet* and VGG19*. The gap of ResNet50* compared to Xception and VGGNet16, is 26% and 8% respectively. So, ResNet50* has the best accuracy and the Xception has the worst accuracy among all counterparts. Similar conclusions can be drawn for other classification metrics.

If the performance of ResNet50 model is compared with the performance of ResNet50* and VGG16*, then the deviation in accuracy is 6% and 4% respectively, and in compare with Xception architecture the accuracy improved 20%. In other words, the accuracy of ResNet50 architecture is 20% higher than that of the Xception. The ResNet50 model proved to be most effective at classifying examples belonging to the classes BCC and NV.

In Table 4, we list the classification performance of all algorithms on different skin lesion classes. ResNet50 had better performance on most of the classes including AKIEC, VASC, NV and MEL and the highest average accuracy and lowest standard deviation of 92.99% and 7.17% respectively.

The feature maps extracted from different CNN layers of a BKL (benign keratosis) class image is shown in Fig. 4. We extract feature maps of some layers for a better demonstration. We can see that, feature maps in the low-level CNN layers is able to preserve the spatial information in order to learn the parameters from these layers but discard that spatial information in the high-level layers and become more abstract. The quantitative obtained results for five architectures are also summarized in the form of confusion matrices in Fig. 5.

Receiver Operating Characteristics (ROC) curves is a graphical way to show the TPR against the FPR obtained by classification thresholds values. Fig. 6. shows Receiver Operating Characteristics (ROC) curves for the five architectures on different skin lesion classes of the ISIC2018 dataset.

Referring to the Fig. 6, it is clearly observed that ResNet50 outperform the other four counterparts with the highest detection recall and the lowest false positive rates. The ROC curve for ResNet50 gains the optimal value one for AKIEC, BCC, BKL, NV and VASC classes. This ensures high detection accuracy of the ResNet50 compared to the other counterparts.

Table 5 Shows the number of errors produced by each model from 11,233 samples of test set. Based on Table 5, we see that ResNet50 and Xception have the lowest misclassified samples.

All these parameters state that the ResNet50 approach performs better in all aspects. This study supported deep learning algorithms with automatically generated features have comparable discriminative power in skin cancer diagnosis on dermoscopic images, and the well-tuned deep learning algorithms such as ResNet50 and VGG16 and Xception have better performance in terms of accuracy (Table 3). The results also show that our proposed preprocessing steps can boost the deep learning algorithm performance compared to original images. To the best of our knowledge, this study is the first reported study compared the performance of five deep learning algorithms on melanoma detection.

5. Conclusion

In this paper, a comparative study is presented for the skin lesion cancer classification using dermoscopic images. Pre-processing methods such as illumination correction, contrast enhancement and artefact removal have been used to improve image quality and obtain a better generalization ability. Due to the imbalance class distributions of skin lesions, we used various augmentation approaches such as horizontal

Table 4
The classification performance of different methods based on different classes.

Methods	AKIEC	BCC	BKL	DF	MEL	NV	VASC	Average ± Std
AlexNet	0.9111	0.8492	0.7053	0.9836	0.80	0.6816	0.9605	0.84 ± .11
VGGNet16	0.9726	0.8826	0.7535	0.9933	0.9438	0.7871	**1**	0.90 ± .0.09
VGGNet19	0.8749	**0.961**	**0.8262**	**1**	0.8764	0.7625	**1**	0.89 ± 0.08
ResNet50	**0.9743**	0.929	0.7717	0.9905	**0.9465**	**0.9232**	**1**	**0.92** ± 0.07
Xception	0.9646	0.914	0.7652	**1**	0.8911	0.8427	**1**	0.90 ± 0.08

The best values under the same conditions are shown boldface.

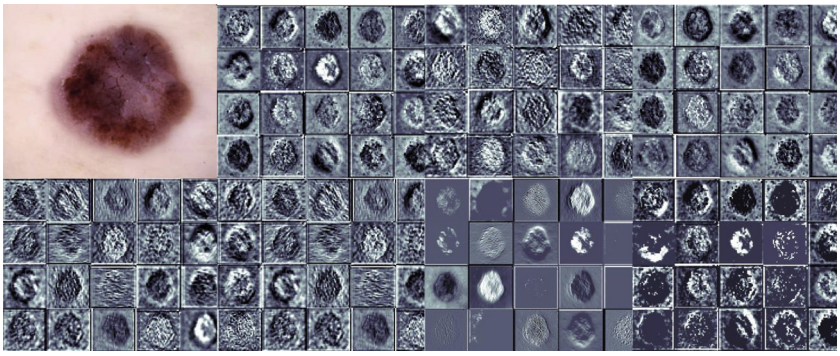


Fig. 4. Feature maps of convolutional layers learned by our ResNet50 model.

and vertical flip, random contrast and random brightness. Augmenting data enabled us to increase the size of training set and reduce the overfitting problem. We employed various standard evaluation metrics such as specificity, sensitivity, accuracy, F-measure to evaluate the obtained results. Our experiments show that ResNet50 outperforms the counterparts AlexNet, Xception, VGGNet16 and VGGNet19 architectures with classification accuracy as high as 92.08% and F-score equal to 92.74%. For future research direction, we aim to boost

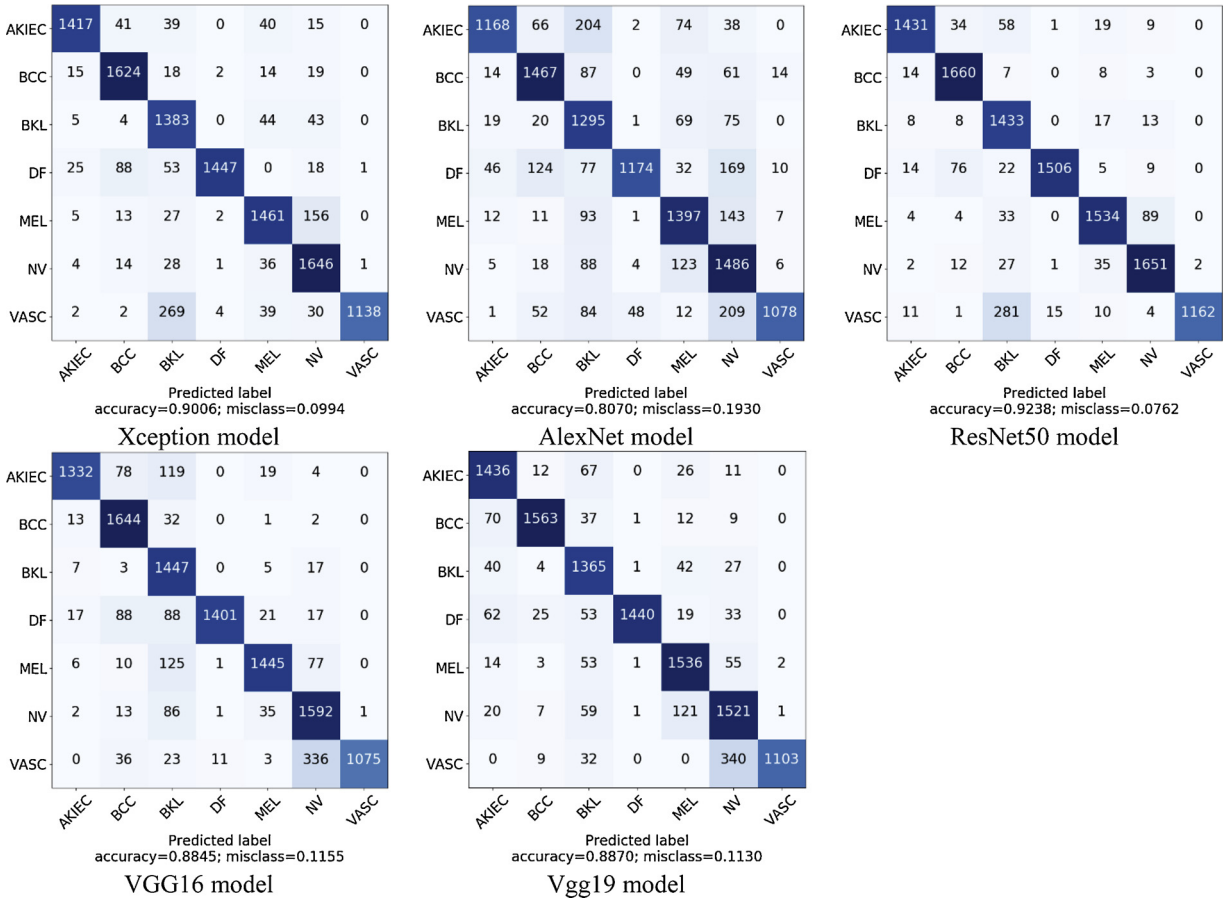


Fig. 5. Confusion matrix of skin lesion classification using different DCNN models. The values on the main diagonal represent all correctly classified instances. The row under each confusion matrix shows the rate of accuracy achieved for each predicted class and misclass predictions.

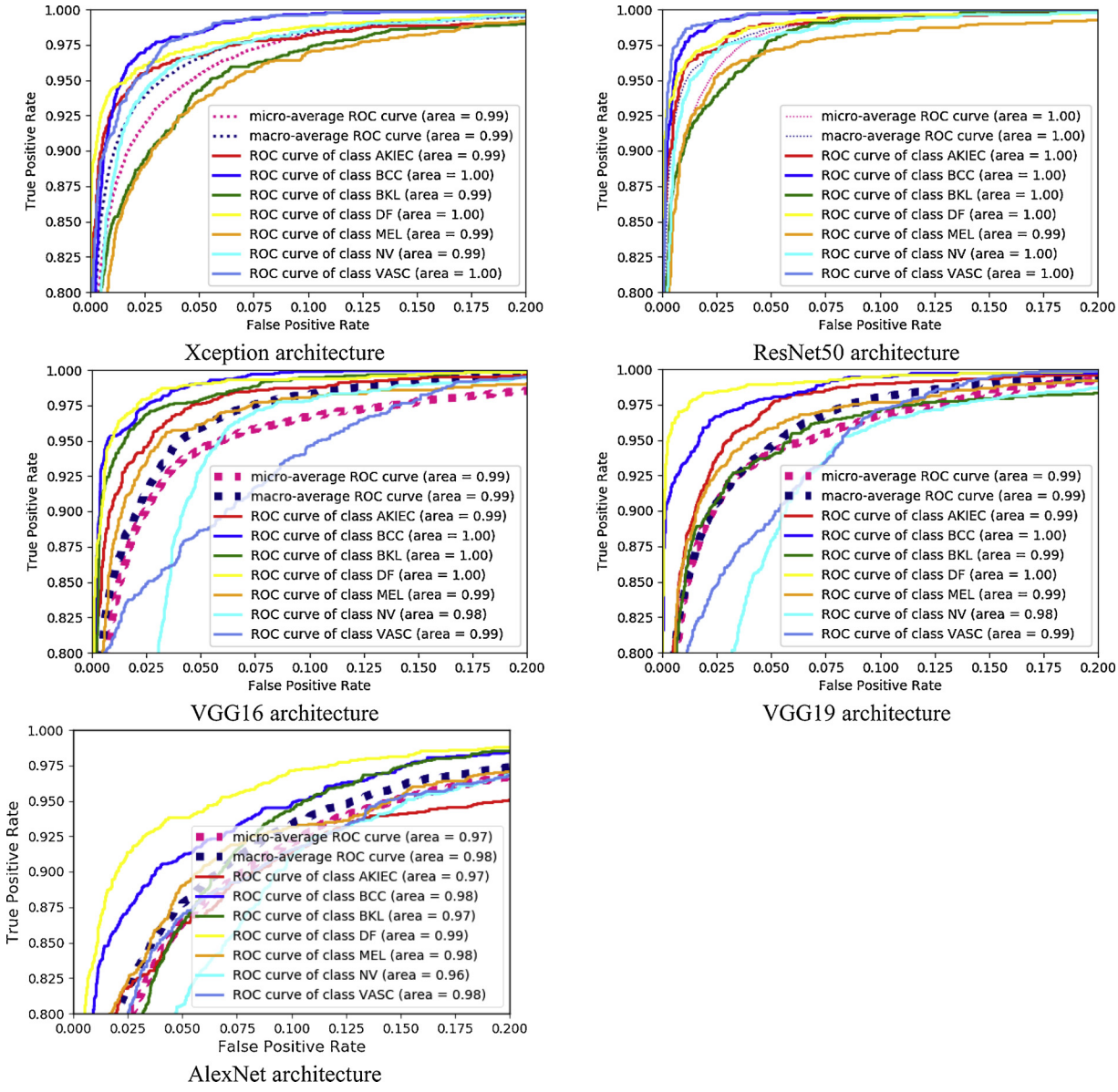


Fig. 6. Comparison of ROC curves of different models.

Table 5
The number of misclassified samples from each model.

	Xception	AlexNet	VGG16	VGG19	ResNet50
Number of misclassified test instances (Total: 11,233 instances)	1139	2092	1305	1272	890

accuracy using deep learning-based ensemble models for melanoma detection.

Disclosure of potential conflict of interest

Non-Declared.

Acknowledgments

The authors would like to express our gratitude to Dr. Ralph Deters of the Department of Computer Science from University of Saskatchewan, Canada for providing his support of this study. We are also thankful to the reviewers in advance for their comments and suggestions.

References

Al-masni, M.A., Al-antari, M.A., Choi, M.T., Han, S.M., Kim, T.S., 2018. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Comput. Methods Programs Biomed.

"American Cancer Society," 2018. [Online]. Available: https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html.

Amoroso, N., et al., 2018. Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge. J. Neurosci. Methods 302, 3–9.

Ayan, E.H.M.Ü., 2018. Data augmentation importance for classification of skin lesions via deep learning. Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).

Barata, C., Marques, J.S., Celebi, M.E., 2014. Improving dermoscopy image analysis using color constancy. Image Processing (ICIP), 2014 IEEE International Conference on. pp. 3527–3531.

Bi, L., Kim, J., Ahn, E., Feng, D., Fulham, M., 2016. Automatic melanoma detection via multi-scale lesion-biased representation and joint reverse classification. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). pp. 1055–1058.

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 Recognit. 85, 78–89.

Cavalcanti, P.G., Scharcanski, J., Lopes, C.B.O., 2010. Shading attenuation in human skin

color images. International Symposium on Visual Computing. pp. 190–198.
Chollet, F., 2017. Xception: deep learning with depthwise separable convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.
Codella, N.C.F., 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). Proceedings - International Symposium on Biomedical Imaging.
Dorj, U.-O., Lee, K.-K., Choi, J.-Y., Lee, M., 2018. The skin cancer classification using deep convolutional neural network. Multimed. Tools Appl. 77 (Apr. (8)), 9909–9924.
Ercal, F., Chawla, A., Stoecker, W.V., Lee, H.-C., Moss, R.H., 1994. Neural network diagnosis of malignant melanoma from color images. IEEE Trans. Biomed. Eng. 41 (no. 9), 837–845.
Erçal, F., Lee, H.-C., Stoecker, W.V., Moss, R.H., **1**, 1999. Skin Cancer Classification Using Hierarchical Neural Networks and Fuzzy Systems.
Fatima, R., Khan, M.Z.A., Dhruve, K.P., 2012. Computer aided multi-parameter extraction system to aid early detection of skin cancer melanoma. Int. J. Comput. Sci. Netw. Secur. 12 (10), 74–86.
Fayek, H.M., Lech, M., Cavedon, L., 2017. Evaluating deep learning architectures for speech emotion recognition. Neural Netw. 92, 60–68.
Flores, E., Scharcanski, J., 2016. Segmentation of melanocytic skin lesions using feature learning and dictionaries. Expert Syst. Appl. 56, 300–309.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., Sun, Q., 2018. Deep learning for image-based cancer detection and diagnosis − a survey. Pattern Recognit. 83, 134–149.
"ISIC 2018 [Online]. Available: https://challenge2018.isic-archive.com/.
Jafari, M.H., et al., 2016. Skin lesion segmentation in clinical images using deep learning. 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 337–342.
Kawahara, J., BenTaieb, A., Hamarneh, G., 2016. Deep features to classify skin lesions. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). pp. 1397–1400.
Kermany, D.S., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. 1 ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst.
Lei, H., et al., 2018. A deeply supervised residual network for HEp-2 cell classification via cross-modal transfer learning. Pattern Recognit. 79, 290–302.
Liao, H., 2016. A deep learning approach to universal skin disease classification. Univ. Rochester Dep. Comput. Sci. CSC.
Lopez, A.R., Giro-i-Nieto, X., Burdick, J., Marques, O., 2017. Skin lesion classification from dermoscopic images using deep learning techniques. 2017 13th IASTED International Conference on Biomedical Engineering (BioMed). pp. 49–54.
Matsunaga, K., Hamada, A., Minagawa, A., Koga, H., 2017. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv Prepr. arXiv1703.03108.
Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F.V., Avila, S., Valle, E., 2017. Knowledge transfer for melanoma screening with deep learning. Proceedings - International Symposium on Biomedical Imaging.
Naylor, P., Laé, M., Reyal, F., Walter, T., 2017. Nuclei segmentation in histopathology images using deep neural networks. Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on. pp. 933–936.
Okur, E., Turkan, M., 2018. A survey on automated melanoma detection. Eng. Appl. Artif. Intell. 73, 50–67.
Oliveira, R.B., Filho, M.E., Ma, Z., Papa, J.P., Pereira, A.S., Tavares, J.M.R.S., 2016a. Computational methods for the image segmentation of pigmented skin lesions: a review. Comput. Methods Programs Biomed.
Oliveira, R.B., Mercedes Filho, E., Ma, Z., Papa, J.P., Pereira, A.S., Tavares, J.M.R.S., 2016b. Computational methods for the image segmentation of pigmented skin lesions: a review. Comput. Methods Programs Biomed. 131, 127–141.
Pathak, A.R., Pandey, M., Rautaray, S., 2018. Application of deep learning for object detection. Procedia Comput. Sci. 132, 1706–1717.
Pathan, S., Prabhu, K.G., Siddalingaswamy, P.C., 2018. Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—a review. Biomed. Signal Process. Control. 39, 237–262.
Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging 35 (5), 1240–1251.
Sainath, T.N., Mohamed, A., Kingsbury, B., Ramabhadran, B., 2013. Deep convolutional neural networks for LVCSR. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8614–8618.
Satheesha, T.Y., Satyanarayana, D., Prasad, M.N.G., Dhruve, K.D., 2017. Melanoma is skin deep: a 3D reconstruction technique for computerized dermoscopic skin lesion classification. IEEE J. Transl. Eng. Heal. Med. 5, 1–17.
Shao, L., Cai, Z., Liu, L., Lu, K., 2017. Performance evaluation of deep feature learning for RGB-D image/video classification. Ann. Pure Appl. Log. 385-386, 266–283.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR vol. abs/1409.1.
Sun, S., Luo, C., Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. Inf. Fusion 36, 10–25.
Vasconcelos, C.N., Vasconcelos, B.N., 2017. Experiments using deep learning for dermoscopy image analysis. Pattern Recognit. Lett.
Wahab, N., Khan, A., Lee, Y.S., 2017. Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. Comput. Biol. Med. 85, 86–97.
Waheed, Z., Waheed, A., Zafar, M., Riaz, F., 2017. An efficient machine learning approach for the detection of melanoma using dermoscopic images. International Conference on Communication, Computing and Digital Systems (C-CODE). pp. 316–319.
Xu, H., Lu, C., Berendt, R., Jha, N., Mandal, M., 2018. Automated analysis and classification of melanocytic tumor on skin whole slide images. Comput. Comput. Med. Imaging Graph 66, 124–134.
Yang, S., Oh, B., Hahm, S., Chung, K.-Y., Lee, B.-U., 2017. Ridge and furrow pattern classification for acral lentiginous melanoma using dermoscopic images. Biomed. Signal Process. Control 32, 90–96.
Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.-A., 2017. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans. Med. Imaging 36 (4), 994–1004.