



Deep CNN and Data Augmentation for Skin Lesion Classification

Tri-Cong Pham^{1,2}, Chi-Mai Luong^{2,3}, Muriel Visani⁴,
and Van-Dung Hoang⁵(✉)

¹ ThuyLoi University, Dong Da, Hanoi, Vietnam
phtcong@gmail.com

² ICTLab, Vietnam Academy of Science and Technology,
University of Science and Technology of Hanoi,
18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

³ Institute of Information Technology, VAST, Hanoi, Vietnam

⁴ Laboratory L3i, University of La Rochelle, La Rochelle, France

⁵ Quang Binh University, Dong Hoi, Quang Binh, Vietnam
zunghv@gmail.com

Abstract. Deep CNN techniques have dramatically become the state of the art in image classification. However, applying high-capacity Deep CNN in medical image analysis has been impeded because of scarcity of labeled data. This study has two primary contributions: first, we propose a classification model to improve performance of classification of skin lesion using Deep CNN and Data Augmentation. Second, we demonstrate the use of image data augmentation for overcoming the problem of data limitation and examine the influence of different number of augmented samples on the performance of different classifiers. The proposed classification system is evaluated using the largest public skin lesion testing dataset, containing 600 testing images, and 6,162 training images. New state-of-the-art performance result is archived with AUC (89.2% vs. 87.4%), AP (73.9% vs. 71.5%), and ACC (89.0% vs. 87.2%). In additional, we explore the influence of each image augmentation on the three classifiers and observe that performance of each classifier is influenced differently by each augmentation and has better results comparing with traditional methods. Thus, it is suggested that the performance of skin cancer classification and medial image classification could be improved further by applying data augmentation.

Keywords: Medical image · Skin cancer · Deep learning · Data augmentation
Melanoma classification

1 Introduction

More than half of cancer diagnoses world-wide are skin cancer [1, 9, 10]. Basically, there are two types of skin cancer called melanoma and non-melanoma. In recent decades, melanoma's incidence and mortality rate has increased dramatically, thus becomes a major problem in public health. However, early detected patients have

higher chance of curing, especially if the cancer is detected in its early stages and removed, the cure rate can be over 90% [1, 7]. Frequently, skin cancer diagnosis is conducted using visual examination of skin lesion images, and then clinical analysis is conducted if there is a suspicion signal. Automated classification of skin lesions using images inspires the development of artificial intelligence-computer vision [13]. Especially, deep convolutional neural networks (Deep CNN) [15] has been achieved a high level of accuracy in image classification with large datasets [14] and very well suited to the problem of melanoma classification. But applying Deep CNN in melanoma recognition is still a challenge due to insufficiency of labeled data.

In this research, we apply Deep CNN [14, 15, 19, 20] which was pre-trained on approximately 1.28 million images from the 2014 ILSVR Challenge [4] for the problem of melanoma classification [5]. Deep CNN is the most successfully applied and the most accurate deep learning architecture in image classification tasks. It comprised of two main parts which are feature extractor and classifier. Images' features are analyzed and summarized by different layers in CNN, features in each layer represent the level of abstraction of the object, the lower-level features are extracted by the previous layer, and the higher-level features are extracted by the following layer of CNN. In the present work, CNN is used to extract features of input image, and then those features are trained and classified by three methods including Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN).

However, the challenge is the limited train skin lesion data, only 6,162 train images, of those only 1,114 (18.08%) images of melanoma when compared to the data collection of more than 1.28 million images of 2014 ILSVR Challenge. With the relatively small data collection, the accuracy rate is affected by overfitting problem [14]. An effective solution to the issue is Data Augmentation and Dropout [14]. Image augmentation artificially enlarges the dataset through different steps of processing or combination of multiple processing, such as random rotation, translation, mirror, scale, crop, shifts, shear and flips [14]. A key concept of image augmentation is that the generated images do not change the semantic meaning of the original image [14]. This technical has been applied in some melanoma classification researches [8–10, 12, 16–18] and achieved state-of-the-art performance at the time of the researches but still need more deeply study to find out pros-cons of this solution. Furthermore, because of imbalanced data (18.08% melanoma vs. 81.92% non-melanoma) its accuracy lost its meaning of accuracy rate like in normal image classification.

Thus, in the article, we propose a solution of combining Data Augmentation, CNN for feature extraction and NN for classification to overcome the lack of labeled data problem and to improve performance of melanoma classification. Furthermore, we explore different types of classifiers such as SVM, RF, NN and their performance influenced by the image deformations. We validate the effectiveness of the algorithms by the area under the curve (AUC), Average Precision (AP), Sensitivity (SEN), Specificity (SPC) and Positive Predictive Value (PPV) instead of Accuracy as a normal classification system.

2 Proposed Classification System

In this study, we propose a melanoma classification system which includes three main components: augmenting data module, extracting features module and classifier as shown in Fig. 1.

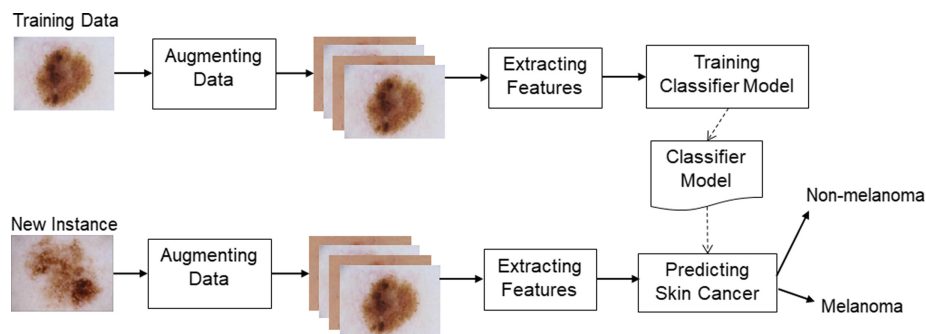


Fig. 1. Proposed classification system.

In the classification system, the data augmentation is applied using traditional transformation of Google’s TensorFlow, the feature extractor is implemented by Deep CNN, the classifier is built by traditional algorithms such as NN, SVM, RF. The order of training process is: augmenting skin lesion image, extracting features of generated images, training classifier model, and saving classifier model to model file. In predicting process, these steps are: augmenting skin lesion image, extracting average feature vector of augmented images, predicting label of the image by the average feature vector and classifier’s model file.

3 Data Augmentation and Deep CNN

3.1 Data Augmentation

Recently, Data Augmentation has been widely being used by not only natural image classification but also melanoma classification such as Matsunaga et al. [16], González-Díaz [12], Menegola et al. [17, 18], Esteva et al. [10], Codella et al. [9]. This is the easiest and most common method to mitigate overfitting problem of scarcity of labeled data in melanoma classification. The most importance concept of image augmentation is that the deformations applied to the annotated data do not change the semantic meaning of the labels. In melanoma classification, we apply three types of data augmentation.

Geometric augmentation: The skin lesion scale and position within the image still maintains the semantic meaning of the lesion, thus does not alter its final classification. Therefore, input images were transformed to generate new samples with the same label of original one by random combination of cropping and horizontal, vertical flips.

Color augmentation: The skin lesion images are collected from different resources and are created by different types of device. Therefore, it is important to normalize the colors of the images when we use them for training and testing any system to improve performance of classification system [6, 11].

Data warping based on specialists knowledge: The fact that the melanoma specialist diagnosis is performed over the observation of the patterns around the lesion. In machine learning, affine transformations such as shearing, distorting and scaling randomly warp stroke data for image classification [21]. Thus warping is very well suited to augment data for improving performance and mitigating overfitting of melanoma classification.

In current research, the data augmentation module combines these three types of augmentation in two steps. Firstly, we normalize input image by adding multiples a converting all pixels into $[-1.0, 1.0]$ range to create normalized data. Secondly, we combine cropping, scaling, distorting and horizontal, vertical flips processes in one step to augment the normalized data. In this step, we apply random parameters of each function to generate samples from original one.

3.2 Feature Extraction

In the classification system shown in Fig. 1, we use Deep CNN as feature extractor. There are many models of CNN such as AlexNet, GoogleNet, ResNet and so on. Our research uses Inception V4 [19] to implement the feature extraction process.

Inception V4 is a well-known architecture developed base on the GoogLeNet platform, this is an upgrade of Inception V3 [20], the input of this network is an image (299×299 pixels), the output depends on how many classes targeted to predict. In the pre-trained model used in this research, the output is the 1000-categories. However, in this study, we use the latest version GoogLeNet (Inception V4); remove full-connected layer and use the Average Pooling layer as the final layer. The output of Average Pooling layer is a 1-Dimension of 1,536 floating numbers. Beside the 1-Dimension features at the Average Pooling layer, three logit features are added to final features from the CNN network, therefore overall of 1-Dimension output is 1,539 floating numbers for each image. The architecture of Inception V4 [19] is demonstrated in Fig. 2.

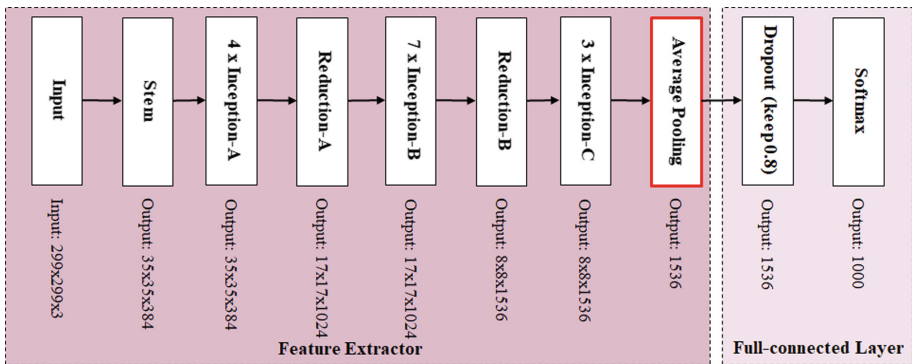


Fig. 2. The overall schema of the Inception-V4 network.

As shown Fig. 2, Inception V4 has two parts, feature extractor and full-connected layer. In detail, the feature extractor has many convolutional blocks include one Stem block, four Inception-A blocks, seven Inception-B blocks, three Inception-C blocks and one Average Pooling layer. The full-connected layer is combined by one dropout block and one softmax layer. Technically, the stem module uses Conv and MaxPool block to convert $299 \times 299 \times 3$ image shape into $35 \times 35 \times 384$ shape which is the input of Inception-A block. In the other hand, Inception-A, Inception-B, Inception-C blocks use only Conv and Avg Pooling to convolute higher abstract features of images. While Inceptions with same type have same shape size and connect directly in sequence, Inceptions with different type need a reduction grid-module to connect together. For instance, Reduction-A grid-reduction module which converts a 35×35 shape to a 17×17 shape is used to connect Inception-A block and Inception-B block. Moreover, Reduction-B grid-reduction module which converts a 17×17 shape to an 8×8 shape is used to connect Inception-B block and Inception-C block. The output ($8 \times 8 \times 1536$ shape) of Inception-C block is converted into 1-Dimension of 1536 features by average pooling layer. In Inception-V4 schema, although the features of average pooling layer is changed by dropout layer (keep 0.8) then trained or classified by softmax as full-connected layer, we save this features to a file as output of feature extractor component in this research as shown in Fig. 1.

3.3 Classifier

In this study, we empirically investigate the influence of data augmentation on three machine learning classifiers, and attempt to answer the question of should we apply data augmentation to improve performance of melanoma classification. The classifiers are trained by the same features which are extracted from augmented images by Deep CNN. In testing process, all of these classifiers are also tested by same test dataset of 2017 ISBI Challenge with three types of data augmentation (without data augmentation - NO DAUG, augments 50 samples each image - DAUG 50, augments 100 samples each image - DAUG 100).

Support Vector Machine (SVM) is a well-known classifier in many image classification, and also showcases outstanding performance [7] in melanoma classification. Therefore, we use LinearSVC library of sklearn-python as SVM model to compare with other classifiers. We do not create new SVM algorithm, we modify the parameters of TOP #3 [18] instead. We find out the best estimators of LinearSVC by hyperparameter tuning with RandomizedSearchCV library of sklearn-python. The best estimators are the same for NO-DAUG, DAUG-50 and DAUG-100 with $C = 0.0003668060861257186$, Dual = False, Multi Class = "ovr" and Random State = 0.

Random Forest (RF) is a popular classifier in machine learning. Thus, we explore the influence of its performance on image augmentation in this study. While the best estimators of DAUG-50 and DAUG-100 are ($n_estimators = 200$, $max_features = "sqrt"$, $max_depth = 10$, $min_samples_split = 5$, $criterion = "entropy"$), they are ($n_estimators = 50$, $max_features = "sqrt"$, $max_depth = 10$, $min_samples_split = 10$, $criterion = "entropy"$) for NO-DAUG.

Neural Network (NN) is a traditional algorithm. In this research we apply a simple NN comprised of two hidden layers to be the classifier with activation using sigmoid in Keras. The loss function is the binary_crossentropy. The parameters of NN are shown in Table 1.

Table 1. NN parameters.

Layer	Parameter	Value	Layer	Parameter	Value
1 st Activation	Type	ReLU	Dropout	Threshold	0.5
1 st Activation	input_shape	(1539)	Optimizer	Type	rmsprop
1 st Activation	Output	256	Optimizer	Metrics	sensitivity, specificity
2 nd Activation	Type	sigmoid	Optimizer	Loss	binary_crossentropy
2 nd Activation	Output	1			

4 Experimental Results

In order to focus exclusively on the improvements obtained by data augmentation, the research implements the classification system in Python, then empirically investigate the influence of data augmentation on three machine learning classifiers, and attempt to improve performance of melanoma classification task with the 2017 ISBI Challenge test dataset.

4.1 Datasets

In this research, we use a data collection of 6,162 train images and 600 testing images which are the same with the test images of ISBI Challenge [5]. Besides, the train dataset includes 2000 images from the ISBI Challenge [5], and the remainder images are collected from the sources such as ISIC Archive [2], PH2 Dataset [3] in Table 2.

Table 2. Data resources.

Resources	Melanoma	Non-melanoma	Total
ISIC 2017 challenge train	374	1,626	2000
ISIC 2017 challenge test	117	483	600
ISIC archive	700	3,262	3,962
PH2 dataset	40	160	200
Total	1,231	5,531	6,762

Table 3 below describes in details the number of images in two types: melanoma and non-melanoma in new train and test datasets. Although we still keep test data of ISBI Challenge and add 4,162 images to the train dataset, the percentage of melanoma images in train dataset is almost same (18.1% in new train dataset vs. 18.7% in train dataset of ISBI Challenge 2017) compare with this percentage in test dataset is 19.5%.

Table 3. Train and test datasets.

Skin lesion	Training	Training %	Testing	Testing %
Melanoma	1,114	18.1%	117	19.5%
Non-melanoma	5,048	81.9%	483	80.5%
Total	6,162	100.0%	600	100.0%

4.2 Evaluation

In this paper, we have chosen to use the ISBI Challenge 2017 [5] dataset for evaluation, which is used by dozens of prior algorithms, in order to archive many result for comparison. The ISBI Challenge provides 2,000 labeled images (374 of melanoma vs. 1,624 of non-melanoma) as train dataset and 600 labeled images (117 of melanoma vs. 483 of non-melanoma) as test dataset. This challenge is the largest standardized and comparative study in this field to date because it has not only the biggest number of training and testing datasets but also 23 finalized test submissions with different algorithms in melanoma classification task. Although the challenge's winner is announced, the train and test datasets are still available for further research and development. In the following, the results of the research are compared with the top ranked participant submissions by the same effectiveness measures of the challenge.

Effectiveness measures: To evaluate the effectiveness of a classification model, we have many effectiveness measures, depending on our data. Our study is using five measures to evaluate classification system such as Area Under the Curve (AUC), Average Precision (AP), Sensitivity (SEN), Specificity (SPC) and Positive Predictive Value (PPV). Mathematically, SEN, SPC and PPV can be expressed base on true positive (TP), true negative (TN), false positive (FP), false negative (FN) as below:

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{SPC} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

In melanoma classification of ISBI Challenge [5], the results are ranked and awarded in AUC. The top three submissions are shown in Table 4 as TOP #1 [16], TOP #2 [12] and TOP #3 [18]. The submissions are state-of-the-art performances in melanoma classification at the time of submission (the highest AUC is 87.4%). This research and the three winners use the same test dataset of the challenge. Besides, while this study uses 4,162 external train images, compares with 1,444 images of TOP #1 and 7,640 images of TOP #3, TOP #2 does not use external training images. Technically, all the winners use Deep CNN, while TOP #1 and TOP #2 use 50-layer ResNet [19] with full-connected layer, TOP #2 uses Deep CNN as feature extractor and SVM as classifier. Our research also uses Deep CNN as feature extractor; and explores performances of three classifiers (NN, SVM, RF) with three types of data augmentation (NO DAUG, DAUG 50 and DAUG 100). The final results are shown in Table 4.

Table 4. Performance of classifiers with same test dataset with 600 images.

Classifier		AUC	AP	SEN	SPC	ACC	PPV
ISBI TOP 3	TOP #1	0.868	0.710	0.735	0.851	0.828	–
	TOP #2	0.856	0.654	0.103	0.998	0.823	–
	TOP #3	0.874	0.715	0.547	0.950	0.872	–
DAUG 100	NN	0.892	0.739	0.556	0.971	0.890	0.823
	SVM	0.773	0.547	0.581	0.965	0.890	0.800
	RF	0.751	0.530	0.530	0.973	0.887	0.827
DAUG 50	NN	0.882	0.736	0.598	0.950	0.882	0.745
	SVM	0.775	0.727	0.590	0.961	0.888	0.784
	RF	0.757	0.526	0.547	0.967	0.885	0.800
NO DAUG	NN	0.862	0.696	0.581	0.942	0.872	0.708
	SVM	0.771	0.522	0.590	0.952	0.882	0.750
	RF	0.746	0.517	0.521	0.971	0.883	0.813

For the first experiments, the combined Data Augmentation and NN solution archived new state-of-the-art performance in melanoma classification task with AUC (89.2% vs. 87.4%), AP (73.9% vs. 71.5%), and ACC (89.0% vs. 87.2%). To get there results, the NN is trained on extracted features of DAUG 100 and run on 500 epochs, 15% validation split (923 validation images vs. 5,239 train images), the binary_crossentropy as loss function, and min loss checkpoint. Although the training is run on 500 epochs, the min loss value is reached (0.08236) in the first epoch. The classification scores is normalized between 0.0 to 1.0 and any confidence above 0.5 is considered positive. In additional, at the same time of writing this paper, the new-state-of-the-art solution is published by Codella et al. [8] performed same AUC (89.2%) with the proposed method. Besides, the results are outperformance when compare with prior algorithms of Gutman et al. [13] (AUC: 89.2% vs. 80.4%, SEN: 55.6% vs. 50.7%, SPC: 97.1% vs. 94.1%, AP: 73.9% vs. 63.7%) and Codella et al. [9] (AUC: 89.2% vs. 84.3%, SPC: 97.1% vs. 83.6%, AP: 73.9% vs. 64.9%).

The next experiments show the influence of each image augmentation on three classifiers. We examine the difference between the effectiveness measures by three proposed classifiers with and without augmentation. According to the result of Table 4, we see that overall the classification performances are improved when applied data augmentation for all classifiers. With NN classifier, the more samples are augmented, the higher effectiveness of all measures is performed; and NN is the best classifier when we compare it with SVM and RF. However, although SVM and RF classifiers performed better accuracy with DAUG 50, we observe that image augmentation can also have a detrimental effect when we increase the number of samples (DAUG 100). For instance, SVM and RF algorithms perform the highest AUC, AP and SEN when applied DAUG 50, but they are affected negatively when data augmentation samples are increased to 100 samples each image.

5 Conclusions

In the study, we propose the use of Data Augmentation and Deep CNN to improve performance of melanoma classification and explore the influence of image augmentation on three classifiers performances. As shown in the experimental results, the two primary contributions of the approach include: (1) new cutting-edge performance is archived with the AUC (89.2%), AP (73.9%), and PPV (82.3%) in melanoma classification task on test dataset of ISIC 2017 Challenge [5]. (2) we examined the influence of skin lesion image deformations on performances of three classifiers (NN, SVM, RF). We observed that the performances of all three classifiers are influenced and improved differently by data augmentation. While NN classifier archived the best performance, SVM and RF classifier's performance had detrimental effect when we changed data augmentation from DAUG 50 to DAUG 100.

However, in this research, although AUC achieved the state-of-the-art performance but SEN is still average comparing to TOP #1 [16] and needs more improvement. Besides, the combination of another architectures or lower layer of Deep CNN and data augmentation is still a challenge for researchers. Furthermore, researchers could still carry out the method of fine-tuning at the last layer to reuse the weight of the network trained by 1.2 million images, because skin lesion images have many similarities with trained natural images.

References

1. American Cancer Society: Cancer facts and figures 2016. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2016/cancer-facts-and-figures-2016.pdf>. Accessed 15 Oct 2017
2. The International Skin Imaging Collaboration (ISIC). <https://isic-archive.com/>. Accessed 15 Oct 2017
3. PH2 Dataset. <https://www.fc.up.pt/addi/ph2%20database.html>. Accessed 15 Oct 2017
4. Large Scale Visual Recognition Challenge 2014 (ILSVRC 2014). <http://image-net.org/challenges/LSVRC/2014/>. Accessed 15 Oct 2017
5. ISIC 2017: Skin Lesion Analysis Towards Melanoma Detection. <http://challenge2017.isic-archive.com>. Accessed 15 Oct 2017
6. Barata, C., Celebi, M.E., Marques, J.S.: Improving dermoscopy image classification using color constancy. *IEEE J. Biomed. Health Inform.* **19**, 1146–1152 (2014)
7. Codella, N.C.F., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R.: Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In: Zhou, L., Wang, L., Wang, Q., Shi, Y. (eds.) *MLMI 2015*. LNCS, vol. 9352, pp. 118–126. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24888-2_15
8. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kallou, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection. In: *A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI)*, Hosted by the International Skin Imaging Collaboration (ISIC). ArXiv e-prints [arXiv:1710.05006](https://arxiv.org/abs/1710.05006) [cs.CV] (2017)

9. Codella, N.C.F., Nguyen, Q.B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., Smith, J. R.: Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM J. Res. Dev.* **61**(4), 5 (2017)
10. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017)
11. Ercal, F., Chawla, A., Stoecker, W.V., Lee, H.C., Moss, R.H.: Neural network diagnosis of malignant melanoma from color images. *IEEE Trans. Biomed. Eng.* **41**, 837–845 (1994)
12. González-Díaz, I.: Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions. ArXiv e-prints: [arXiv:1703.01976](https://arxiv.org/abs/1703.01976) [cs.CV] (2017)
13. Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection. In: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, Hosted by the International Skin Imaging Collaboration (ISIC) (2016). ArXiv e-prints: [arXiv:1605.01397](https://arxiv.org/abs/1605.01397) [cs.CV]
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NIPS)*, vol. 25, pp. 1097–1105 (2012)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015)
16. Matsunaga, K., Hamada, A., Minagawa, A., Koga, H.: Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. ArXiv e-prints [arXiv:1703.03108](https://arxiv.org/abs/1703.03108) [cs.CV] (2017)
17. Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F.V., Avila, S., Valle, E.: Knowledge transfer for melanoma screening with deep learning. ArXiv e-prints [arXiv:1703.07479](https://arxiv.org/abs/1703.07479) [cs.CV] (2017)
18. Menegola, A., Tavares, J., Fornaciali, M., Li, L.T., Avila, S., Valle, E.: RECOD titans at ISIC challenge 2017. ArXiv e-prints [arXiv:1703.04819](https://arxiv.org/abs/1703.04819) [cs.CV] (2017)
19. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: *Artificial Intelligence*, pp. 4278–4284 (2017)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Computer Vision and Pattern Recognition (CVPR)*, vol. 2016, pp. 2818–2826 (2016)
21. Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D.: Understanding data augmentation for classification: when to warp? In: *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6 (2016)