Skin Lesion Classification using Deep Learning Architectures

Abhishek C. Salian
Dept. of EXTC
SFIT, Borivali
Mumbai, India
abhishekcs26@gmail.com

Shalaka Vaze
Dept. of EXTC
SFIT, Borivali
Mumbai, India
shalakavaze1810@gmail.com

Pragya Singh
Dept. of EXTC
SFIT, Borivali
Mumbai, India
singh03pragya@gmail.com

Gulam Nasir Shaikh
Dept. of EXTC
SFIT, Borivali
Mumbai, India
sgulamnasir346@gmail.com

Santosh Chapaneri
Dept. of EXTC
SFIT, Borivali
Mumbai, India
santoshchapaneri@sfit.ac.in

Deepak Jayaswal
Dept. of EXTC
SFIT, Borivali
Mumbai, India
djjayaswal@sfit.ac.in

***Abstract*—Skin cancer is one of the major types of cancers that can arise from various dermatological disorders and can be classified into various types according to texture, structure, color and other morphological features. Identifying the lesions from skin images can be an important step in pre-diagnosis to aid the doctors and infer the medical condition of the patient. Recent work has focused on classifying only melanoma from a given set of skin lesion images. However, some types of skin lesions (Acctinic Keratosis and basal cell carcinoma) can become malignant over a period of time. So by detecting these classes we can say we are cutting down the risk of malignancy and doing the task of early detection. We are able to classify different types of skin lesions (basal cell carcinoma, benign keratosis, dermatofibroma, vascular lesions, melanoma, and melanocytic nevi) with an accuracy of above 80% with Mobilenet, VGG-16 and our custom model which we have designed. With the help of this models, which will be embedded in skin lesion analyzer machines. This can give the patients as well as doctors a good idea of whether or not there is a need for medical attention and can avoid unnecessary panic/false alarms. We are using different deep learning architectures to classify skin lesions with good accuracy relative to existing work.**

***Keywords*—Skin cancer; Morphological features; Deep learning; Data augmentation**

I. Introduction

Skin cancer is one of the major types of cancers and its incidence has been increasing over the past decades. Deep learning architectures can help us to avoid the step of manual feature extraction. This can save time and can alert the patient if there is a suspicious signal. We have attempted to build a robust and accurate deep learning model that will assist dermatologists in detecting skin cancer and will help to take necessary actions without much delay. By feeding the trained deep learning models with skin lesion image data, the doctor can know the type of lesion and decide whether it holds the potential to metastasize in the future or not. There are higher chances of curing, if the cancer is detected in its early stages, the cure rate can be about over 90% [2].

Skin cancer diagnosis is conducted using visual examination of the lesion and then the clinical analysis is conducted if there is a suspicion. Image-based classification using deep learning, in particular, have recently shown considerable accuracy in medical image classification.

II. Related Work

Existing techniques to solve the given problem are:

1) Using the segmentation network [3]:
   This technique uses a segmentation network that helped to achieved semantic image segmentation to generate an accurate segmentation map of the lesion and to identify the lesion that corresponds in the image. The network extracts the lesion part from the image to broadly define the lesion which gives a binary map. Further, this lesion mask is given to an augmentation network where the image is augmented by applying various functions. The augmented images are further given to structure segmentation network to provide a corresponding segmentation into a pre-defined set of textural patterns and local structures that are of special interest for dermatologists in their diagnosis. In particular, the network recognizes these set of eight structures:
   a) Dots, globules, and cobblestone pattern
   b) Reticular patterns and pigmented network
   c) Homogenous areas
   d) Regression areas
   e) Blue-whitish veil
   f) Streaks
   g) Vascular structure
   h) Unspecified pattern
   These are given to the diagnosis network which consist of ResNet-50 architecture for classifying two classes melanoma and non-melanoma.
2) Recognition of melanoma by using sparse coding [4]: It uses two parallel paths:

978-1-7281-2339-4/20/$31.00 ©2020 IEEE

a) Transfer of convolutional neural network features learned from the domain of natural photographs

b) Unsupervised feature learning, using sparse coding, within the domain of dermoscopy images.

Classifiers are then subsequently trained for each using non-linear SVMs, and the models are then combined in late fusion (score averaging). The Caffe Convolutional Neural Network (CNN) was used for transfer learning. The sparse coding algorithm helped to eliminate the need for large collections of annotated data to learn good features, and allowing the system to draw analogies. Augmentation was done on the dataset to improve the performance and fusion of both i.e. sparse coding and Caffe Convolutional Neural Network helped to achieve the expected accuracy.

3) Image Classification of Melanoma, Nevus and Seborrhoeic Keratosis by Deep Neural Network Ensemble (2017) [5]:
To classify skin lesion images into three classes – melanoma (MM malignant melanoma), nevus (NCN; nevocellular nevus) and seborrheic keratosis (SK) through two binary classifiers – MM vs. rest (MM classifier) and SK vs. rest (SK classifier). The mentioned paper makes use of external training data and the use of the age/sex information tagged with a number of the provided samples. The luminance and colour balance of input images are normalized exploiting colour constancy. Normalized images are input to a base classifier trained for SK vs. rest as well as to a base classifier trained for MM vs. rest. Both base classifiers have identical composition. Geometrically transformed images (combinations of rotation, translation, scaling and flipping) are input in parallel to an ensemble of convolutional neural networks (CNNs) and a prediction value in [0.0, 1.0] is output. Adoption of a 50-layer ResNet implemented with small modifications was used as the architecture. A straightforward thresholding was adopted by age/sex information only for SK classification. For MM classification, it was observed no significant increase by cross-validation. In addition, it was noticed that SK classifier was far more reliable than MM classifier. Ad-hoc linear approximation was applied. CNNs were fine-tuned with the training samples from the initial pre-trained model for generic object recognition in Keras. They applied different types of optimization and selected the best combination of fine-tuned CNNs through cross-validations. The optimization methods used were RMSProp and AdaGrad.

III. Proposed Classification System

In this paper, we propose a skin lesion classification which includes augmenting the labeled images, extracting the features and predicting the skin lesion.

The classification system shown in Fig. 1 takes an RGB image as an input, size of input image depends upon deep learning architecture used, for example 224 × 224 × 3 is used as an input image size for MobileNet architecture, Custom Model and VGG-16, etc.

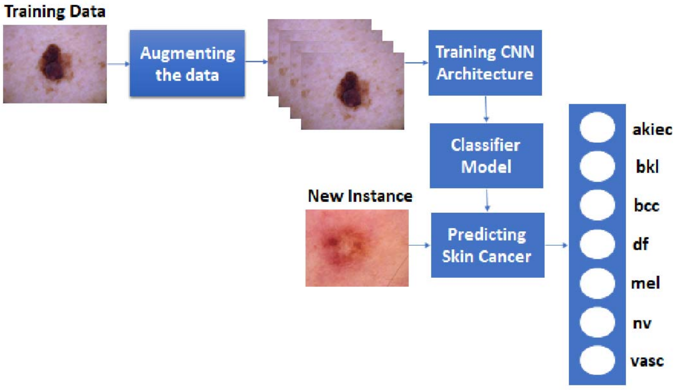

Fig. 1. Proposed Classification System

The training process is: augmenting the labeled skin lesion images as explained in III-B, the augmented images are then fed to the CNN Architecture. After training is done, the trained model is used for predicting the type of skin cancer for any new instance. The trained model which gives minimum validation loss while training is been used, for example while training the MobileNet model, suppose we get minimum validation loss at 11^{th} epoch while training, then that model is being saved using a Keras callback method and that model is being used for evaluating its performance amongst other deep learning architecture models such as VGG-16, etc.

The predicting process is: the best model which is saved during the training process will be loaded from a file, the new instance of skin lesion image data is given as an input to this deep learning classifier model. The output of the classifier model will give a prediction on the type of skin lesion identified from the given input sample image.

A. Datasets

1) PH^2 Dataset [6]:

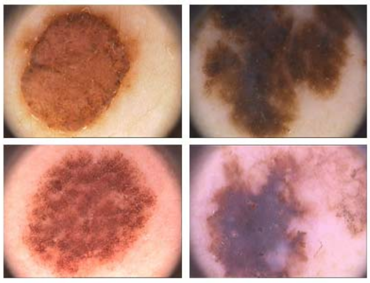

Fig. 2. Dermoscopic images from the PH2 dataset

PH^2 is a dermoscopic image dataset, some of the sample images of this dataset are shown in Fig. 2. Each image in the PH^2 dataset is classified into either non-melanoma (common nevus and atypical nevus) or melanoma. It consists of a total number of 200 lesion images, including 80 common nevi, 80 atypical nevi, and 40 melanoma

TABLE I
PH^2 DATASET WITHOUT AUGMENTATION

Classes	Non-Melanoma		Melanoma
	Atypical Nevi	Common Nevi	
Total Images	80	80	40

images as shown in Table I. Here, common nevi and atypical nevi can be combined to be termed as non-melanoma which consist of 160 images. Due to limited number of samples in the dataset, validating and relying on the results obtained using this samples cannot be generalized well by different machine learning algorithms and deep learning models. This dataset consists of RGB color images with a maximum resolution of 768×560 pixels and there are many samples in dataset with different resolutions. All samples in this dataset was evaluated by a dermatologist concerning the following parameters:

a) Manual segmentation of the skin lesion
b) Clinical and histological diagnosis
c) Dermoscopic criteria (Asymmetry, Colors, Pigment network, Dots/Globules, Streaks, Regression areas, Blue-whitish veil)

The lesions can be divided into two main groups considering their nature:

a) Benign lesions (which include common Atypical nevus)
b) Malignant lesions (melanoma).

TABLE II
PH^2 DATASET WITH AUGMENTATION.

Classses	Non-Melanoma	Melanoma
Total Images	2000	2000

The labeled images of the PH^2 dataset was augmented using Augmentor refer III-B, to get a total of 4000 labeled images. It consists of 2000 images of Non-Melanoma and 2000 images Melanoma, which forms a balanced dataset for training deep learning models as shown in Table II.

2) HAM10000 Dataset [7]:
The following are the description of diagnostic categories:

TABLE III
HAM10000 DATASET MANUALLY AUGMENTED

Total Images	akiec	bcc	bkl	df	mel	nv	vasc
10015	327	514	1099	115	1113	6705	142

The problem of small size and lack of available dataset of dermatoscopic images is tackled using HAM10000 dataset. The publically available HAM10000 dataset consists of 10015 labeled dermoscopic images which consist of manually augmented images. Here the manual augmentation was done by cropping the images with lesions in the centre, changing the magnification size, changing the histogram, etc. This dataset is divided into seven categories of images such as 327 images of Actinic Keratoses (akiec), 514 images of Basal Cell Carcinoma (bcc), 1099 images of Pigmented Benign Keratoses (bkl), 115 images of Dermatofibroma (df), 1113 images of Melanoma (mel), 6705 images of Nevus (nv) and 142 images of Vascular (vasc). This dataset was pathologically verified by the dermatologists refer Table III.

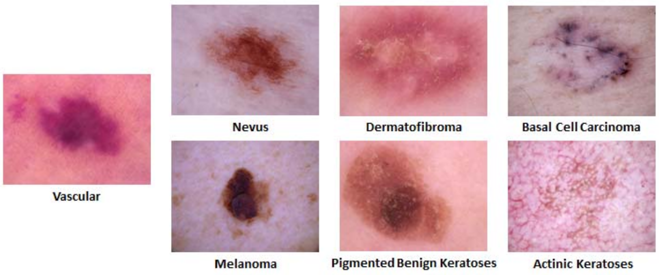

Fig. 3. Disease classification within dermoscopic images

TABLE IV
HAM10000 DATASET WITHOUT AUGMENTATION

Total Images	akiec	bcc	bkl	df	mel	nv	vasc
5515	151	175	440	39	230	4415	64

The HAM10000 dataset without augmentation consists of 5515 dermoscopic images which consist of images without augmentation. This dataset are divided into seven categories of images such as 151 images of Actinic Keratoses, 175 images of Basal Cell Carcinoma, 440 images of Pigmented Benign Keratoses, 39 images of Dermatofibroma, 230 images of Melanoma, 4415 images of Nevus and 64 images of Vascular. This dataset was pathologically verified by the dermatologists refer Table IV.

TABLE V
HAM10000 DATASET AUGMENTED BY US USING AUGMENTOR

Total Images	akiec	bcc	bkl	df	mel	nv	vasc
34415	5000	5000	5000	5000	5000	4415	5000

The labeled images of the HAM10000 dataset was augmented using Augmentor to a total of 34415 labeled images to balance each category of skin lesion. This dataset consists of images such as 5000 images of Actinic Keratoses, 5000 images of Basal Cell Carcinoma, 5000 images of Pigmented Benign Keratoses, 5000 images of Dermatofibroma, 5000 images of Melanoma, 4415 images of Nevus and 5000 images of Vascular Refer Table V.

B. Data Augmentation

Data augmentation is useful for creating more samples out of limited number of samples contained in original dataset. This augmented dataset consist images of different orientation. HAM10000 and PH^2 dataset consist of unbalanced images for each class shown in Table III which does not provide expected performance. Therefore, in order to train the models, we require a balanced dataset to reduce biasness.

Augmentor Library: Augmentor is a Python library used for augmentation of image samples. We have used Augmentor for augmenting images which consists of operations such as elastic distortion, rotation, shearing, cropping, mirroring and skewing. The chosen augmentation operations used in our project include rotation, flip left and right, flip top and bottom. Augmentor applies operation to the images in stochastic manner according to user defined probability value for each of the operation mentioned above. It has outstanding features like size-preserving rotations, size-preserving shearing, and cropping, which is more suitable for machine learning and deep learning model training.

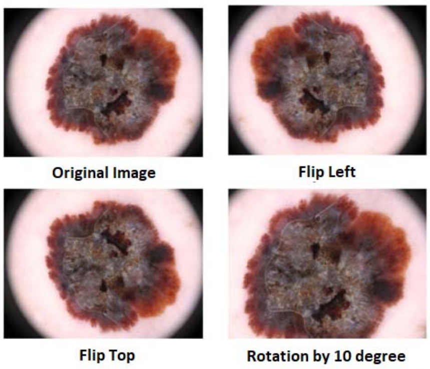

Fig. 4. Image Augmentation

1) Rotation:
   This will rotate by arbitrary degrees, then a crop is taken from the centre of the newly rotated image. We have rotated image by an arbitrary angle in range -10 to 10 degree with a probability of 0.5 which signifies 50% of the sample images from the dataset will be rotated randomly. Zoom effect is not particularly drastic for smaller rotations of between -10 and 10 degrees.
2) Flip (Mirroring):
   a) Flip top and bottom: In this operation, the image will be flipped along the horizontal axis. We have used probability of 0.5 which signifies 50% of the samples will be flipped randomly.
   b) Flip left and right: In this operation, the image will be flipped along the vertical axis. Here also probabilty of 0.5 is used for flipping the samples of dataset.

C. Deep Architectures

We have used several pre-trained CNN architectures as well as our designed custom CNN model as shown in Table VI. The state of the art architectures used are MobileNet and VGG-16 which are pre-trained on Imagenet dataset. The results of all the individual architectures including custom model are shown in section IV.

In the pre-trained models, the output is 1000 categories for all the architectures. By adding a dense layer of 7 neurons, we got the input test images classified into 7 classes (For HAM10000 dataset). In case of PH^2 the same layer consists of 2 neurons.

The input image size for VGG-16, Custom Model and MobileNet architectures is a $224 \times 224 \times 3$ RGB image. The activation function used throughout the architectures including the custom model is ReLU and the optimizer function used is Adam. The loss function used is categorical loss entropy for all the architectures on HAM10000 dataset and binary cross entropy for PH^2 dataset.

TABLE VI
COMPARISON BETWEEN DEEP CNN ARCHITECTURES

Parameters	VGG-16	MobileNet	Custom Model
No. of Layers	16	86	5
Input Image Size	224×224×3	224×224×3	224×224×3
No. of Fully Connected Layers	2	1	2
No. of Maxpool Layers	5	NA	2
No. of Convolutional Layers	14	62	3
No. of Dropout Layers	13	1	3
No. of Batch Normalization Layers	NA	26	NA
Global Average Pooling Used	NA	1	NA

IV. RESULTS

In this paper, we have chosen the publically available and augmented labeled images of PH^2 and HAM10000 dataset for evaluation, which was used by different architectures such as MobileNet, VGG-16 and Custom model to achieve results for comparisons. The PH2 dataset provides 3200 labeled images (1600 of melanoma vs. 1600 of non-melanoma) as a training dataset, 400 labeled images (200 of melanoma vs. 200 of non-melanoma) for validation and 400 labeled images (200 of melanoma vs. 200 of non-melanoma) for test dataset. Whereas the augmented HAM10000 dataset provides 34411 labeled images (34411 labeled images is divided into seven different classes as mentioned in the subsection of the HAM10000 dataset) for training the model, 552 labeled images for validating the model and 551 labeled images for testing the dataset.

For the melanoma classification, the results were evaluated based on Test Accuracy, Area Under the Curve (AUC) and F1-Score for 100 epochs, optimizer as Adam and loss as Cross-entropy loss function.

TABLE VII
RESULTS OF PH^2 DATASET WITHOUT AUGMENTATION

Parameters	MobileNet	Custom Model	VGG-16
Test Accuracy	90%	90%	80%
AUC	0.53125	0.71	0.5
F1 Score	0.9375	0.875	0.8888

TABLE VIII
RESULTS OF PH^2 DATASET WITH AUGMENTATION

Parameters	MobileNet	Custom Model	VGG-16
Test Accuracy	90%	97.25%	50%
AUC	0.61	0.5275	0.5
F1 Score	0.51	0.5141	0.6666

1) For the publically available labeled images of PH^2 dataset achieved 90% Test Accuracy, 0.53125 AUC, 0.9375 F1-Score on MobileNet Architecture; 90% Test Accuracy, 0.71 AUC, 0.875 F1-Score on Custom Model Architecture and 80% Test Accuracy, 0.5 AUC, 0.8888 F1-Score on VGG-16 Architecture refer Table VII.
2) For the augmented labeled images of PH^2 dataset achieved 90% Test Accuracy, 0.61 AUC, 0.51 F1-Score on MobileNet Architecture; 97.25% Test Accuracy, 0.5275 AUC, 0.5141 F1-Score on Custom Model Architecture and 50% Test Accuracy, 0.5 AUC, 0.6666 F1-Score on VGG-16 Architecture refer Table VIII.
3) For the publically available labeled images of HAM10000 dataset achieved 81.52% Test Accuracy, 0.5 AUC, 0.7826 F1-Score on MobileNet Architecture; 83.152% Test Accuracy, 0.5 AUC, 0.668 F1-Score on Custom Model Architecture and 80.07% Test Accuracy, 0.5 AUC, 0.8 F1-Score on VGG-16 Architecture refer Table IX.
4) For the augmented labeled images of HAM10000 dataset achieved 82% Test Accuracy, 0.5 AUC, 0.71 F1-Score on MobileNet Architecture; 80.61% Test Accuracy, 0.509 AUC, 0.76 F1-Score on Custom Model Architecture and 79.71% Test Accuracy, 0.5 AUC, 0.597 F1-Score on VGG-16 Architecture refer Table X.

TABLE IX
RESULTS OF HAM10000 DATASET WITHOUT AUGMENTATION

Parameters	MobileNet	Custom Model	VGG-16
Test Accuracy	81.52%	83.152%	80.07%
AUC	0.5	0.5	0.5
F1 Score	0.7826	0.668	0.8

V. CONCLUSION

In the study, we have used two pre-trained state of the art model i.e. Mobilenet and VGG16, we have evaluated its performance for PH^2 dataset and HAM10000 dataset under two cases, one with data augmentation and the other one without data augmentation. We have also designed our own custom deep learning architecture and compared the performance with the other two models to show that training a well-designed model from scratch can equally perform well. Out of the three models, Mobilenet and custom model performed quite well in terms of its test accuracy, AUC and F1 score evaluation metrics. Data augmentation has no significant effect when compared with without data augmentation results of classifiers. For PH^2 dataset the best model evaluated are the state of art architectures MobileNet and Custom model, as the accuracy for dataset with augmentation and without augmentation is comparably high with respect to VGG-16 as shown in Table VII, VIII.The accuracy achieved for MobileNet is 90% for both, with and without augmentation. In PH^2 dataset, the performance of Custom model classifier is quite good and accuracy evaluated is 97.25% for dataset with augmentation and 90% without augmentation

TABLE X
RESULTS OF HAM10000 DATASET WITH AUGMENTATION

Parameters	MobileNet	Custom Model	VGG-16
Test Accuracy	82%	80.61%	79.71%
AUC	0.5	0.509	0.5
F1 Score	0.71	0.76	0.597

Mobilenet and Custom model performance in the HAM10000 dataset is quite good, refer Table IX, X. The accuracy achieved by the Custom model is 83.152% without augmentation and 80.61% with augmentation while we were able to achieve comparative performance with MobileNet as accuracy is 81.52% without augmentation and 82% with augmentation. These accuracy values also indicate how data augmentation has a detrimental effect on model performance. The performance of both these architectures is good as compared with VGG-16 architecture. The data samples need magnification of 20x which can be achieved by specific tool (tubinger mole analyser) sets the limitation in analysing the lesion. Future work will focus on increasing the accuracy of model and using the proposed model in form of mobile based vision application to identify the specific lesion patterns that may be an indicative of cancer, in order to provide human verifiable evidence to support the disease diagnosis.

REFERENCES

[1] T.-C. Pham, C.-M. Luong, M. Visani, and V.-D. Hoang, “Deep CNN and Data Augmentation for Skin Lesion Classification,” *Intell. Inf. Database Syst.*, pp. 573–582, 2018.

[2] “Skin Cancer Facts & Statistics,” The Skin Cancer Foundation. [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/. [Accessed: 30- Sept- 2019].

[3] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and V. Eduardo, “Knowledge Transfer for Melanoma Screening with Deep Learning,” *IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 297-300, 2017.

[4] N. Codella1, J. Cai1, M. Abedini, and R. Garnavi, “Deep Learning, Sparse Coding, and SVM for Melanoma Recognition in Dermoscopy Images,” *Lect. Notes Comput. Sci.*, vol. 9352, pp. 118–126, 2015.

[5] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, “Image Classification of Melanoma , Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble,” *Comput. Vis. Pattern Recognit.*, pp. 2–5, 2017.

[6] “PH^2 Database”, 2020. [Online]. Available: https://www.fc.up.pt/addi/ph2%20database.html. [Accessed: 30-Jun- 2019].

[7] P. Tschandl, C. Rosendahl, and H. Kittler, “Data Descriptor: The HAM 10000 dataset , a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Nat. Publ. Gr.*, pp. 1–9, 2018.