2020 IEEE International Conference on Advent Trends inMultidisciplinary Research and Innovation (ICATMRI)

Skin Cancer Classification using Transfer Learning

Hari Kishan Kondaveeti
School of Computer Science and Engineering
VIT-AP University
Andhra Pradesh, India
kishan.kondaveeti@vitap.ac.in

Prabhat Edupuganti
School of Computer Science and Engineering
VIT-AP University
Andhra Pradesh, India
prabhatedupuganti@gmail.com

2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI) | 978-1-7281-7734-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICATMRI51801.2020.9398388

***Abstract*— Today, Cancer is one of the major lethal diseases in the world. Globally out of every three cancers diagnosed, one is identified as skin cancer. Some reports suggest that one out of every five Americans might fall prey to skin cancer in the course of their life. Early detection of the disease plays a pivotal role in the treatment of skin cancer. Though these skin lesions can be seen without the help of any external clinical device, it is a challenging task to distinguish between malignant and benign skin lesions as they are alike in their physical appearances. This leads to an increased number of unnecessary biopsies where in one study it was revealed that nearly 5,00,000 biopsies are done in children every year to diagnose a mere 400 melanomas. To tackle this problem and help dermatologists in the diagnosis process, we developed an enhanced image classification model which can act as a preliminary check before moving to a costlier biopsy. This model can identify 7 distinct types of skin lesions. An analysis has been carried out on the HAM10000 dataset. We used transfer learning utilizing multiple pre-trained models, combined with class-weighted loss and data augmentation techniques for the classification process. Experimental analysis shows that the modified ResNet50 model is capable of identifying skin lesion images into one of the seven classes with categorical accuracy, weighted average precision, and recall of 90 percent, 0.89, 0.90, respectively. Our model can be used as a clinical decision support system to help dermatologists in the diagnosis process.**

Keywords*— *Convolutional Neural Networks (CNN), Deep Learning, Transfer Learning, Skin Cancer classification

I. INTRODUCTION

According to the National Cancer Institute, the number of new cancer cases might increase to a staggering 29.5 million and deaths related to the same to 16.4 million per year by 2040 [1]. Exposure to sun and ultraviolet (UV) radiation are major factors for an increase in non-melanoma skin cancers. The magnitude of the above factors is substantial and depends on the amount of exposure [2]. BCC (basal cell carcinoma) which is a non-melanoma is the most common type of skin cancer. The frequency of both non-melanoma and melanoma skin tumors has been expanding over the previous many years. At present, around 2 to 3 million skin malignancies of the former type and 132,000 melanoma skin tumors happen all around the world every year [3].

Melanoma, when detected in the initial stages, is easily curable and the 5-year endurance rate is approximately 98%. Previously surgery is the usual treatment, researchers in this field are now working on other forms of treatment like targeted therapies and immunotherapies [4]. Though skin cancer can be easily detected with an early diagnosis through a basic visual inspection, most patients come under the supervision of doctors when the disease is at advanced stages. In clinical trials, it has been reported that training in dermoscopy prompted enhancements in the early recognition of melanoma for both non-expert dermatologists and primary care doctors. Dedicated clinical solutions have been developed for dermatologists for image acquisition, maintenance as well as retrieval for further follow-up and monitoring [5]. Dermoscopy is a non-invasive indicative procedure to assess pigmented and non-pigmented skin lesions which may not be detectable by assessment with the unaided eye. When dermoscopy is performed by dermatologists without the proper experience, it leads to a reduction in the accuracy of the diagnosis which in turn results in wastage of resources [6].

Instead, the classification of dermoscopic skin lesion patches can be seen as an image classification problem. In recent times, deep learning is being used for medical imaging and has shown remarkable results in various tasks like segmentation, detection, and classification. Despite deep learning being a heavy computation and memory dependent system, it is being viewed as a technology that has unparalleled benefits. It has significantly improved the medical diagnosis workflow. Convolutional neural networks are an improvised version of neural networks initially made for image-related tasks where instead of only fully-connected layers, different filters have been used. Convolutional neural networks work very well for feature extraction but the only problem is that they are data-hungry for training. Transfer learning is a suitable solution to overcome this problem.

In this paper, we developed an image classification model as a solution to the challenging task of a skin cancer diagnosis. We have used convolutional neural networks pre-trained on ImageNet as a fixed feature extractor and add custom layers on the top for our specific task which is a type of transfer learning. In simple words, it is an approach where we used initialized weights of pre-trained networks trained on other data resources. We trained our model on the HAM10000 dataset and it classifies skin lesion images into one of the seven classes present in the dataset. We used ResNet50 [7], InceptionV3[8], MobileNet [9], and Xception [10] as the pre-trained base models. While ResNet50s pre-trained on ImageNet is the current industry standard for computer vision applications, a particularly new technique called BigTransfer (BiT) [11] might become the state-of-the-art transfer learning technique for computer vision in the coming months. This technique has mainly focused on the concept of larger pre-training. This work suggests that BigTransfer performs comparatively well even with a little amount of data.

Section I introduces the concepts related to Skin Cancer and Deep Learning architectures used for Skin Cancer classification in literature. Proposed methodology is described in Section II. Section III and IV presents evaluation metrics used for experimental analysis and results of experimental analysis respectively. The conclusive remarks are presented in Section V.

978-1-7281-7734-2/20/$31.00 ©2020 IEEE

II. PROPOSED METHODOLOGY

A. Dataset

The lack of proper dermatoscopic data in terms of size and diversity has been a setback for training neural networks. The dataset used here is HAM10000 ("Human Against Machine with 10000 training images") [12] dataset released by the Harvard Dataverse. The dataset consists of 10015 dermatoscopic images with a 600x450 pixel resolution. It is a collection of 7 distinct classes. Actinic Keratosis (AKIEC) is benign but potentially malignant and can turn into a malignant lesion. Basal cell carcinoma (BCC) and melanoma (MEL) are malignant. Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Skin Lesion (VASC), and Melanocytic Nevi (NV) are benign. Fig. 1 depicts the images of these classes

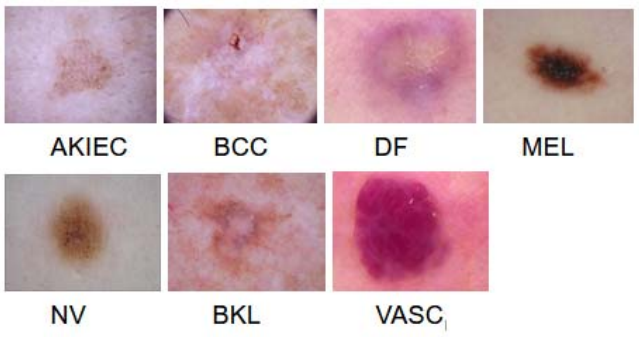

Fig. 1. Different classes in the dataset (Types of skin lesions)

Half of the dataset was confirmed by histopathology. Few other techniques to confirm ground truth were reflectance confocal microscopy, follow-up, or by expert consensus. On examination of the metadata, it has been found that there are only 7470 distinct skin lesions out of the 10015 images. Also, the dataset is extremely imbalanced as depicted in Fig. 2. NV being the largest class has 6705 images while DF, the smallest class has only 115 images.

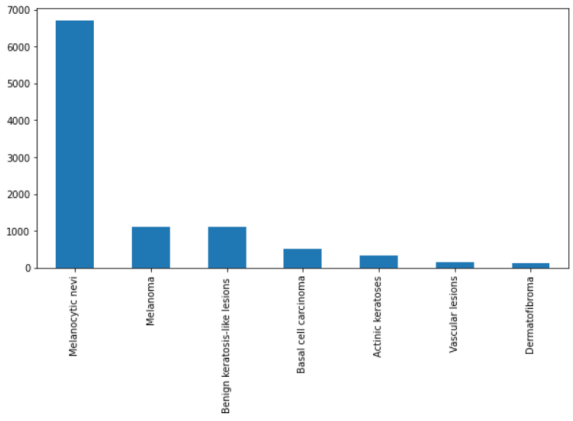

Fig. 2. Class frequency of the dataset

B. Data Preparation

After an initial exploratory data analysis, duplicate images have been identified. After the removal of duplicate images, a validation set has been made with a split ratio of 83:17. After this process, the training set has 9077 images and the validation set has 938 images. Later data augmentation in the form of flipping, cropping, and rotating have been applied with the final target image size set to 224x224 pixels. Stratified sampling has been used to maintain the inter-class ratio. The training set images are later passed through a unique pre-process function concerning the model architecture used. Each architecture requires input to the model in a certain format and the pre-process function helps us transform our data into such required format.

C. Model Architecture and Training

Transfer learning [13] has been used as a technique to train neural networks as they are data-hungry and we only have limited data despite having additional data through data augmentation. After taking this architecture as a base model, the top layer has been removed and a Global average pooling layer, dropout, and dense layers have been added. Different thresholds of dropout have been tried in this study. The training has been implemented on Google Colab using an NVIDIA Tesla T4 provided. Four different architectures namely ResNet50, InceptionV3, Xception, and MobileNet have been used. Class weights have been computed based on the individual class frequency to tackle class imbalance. While training, the weights of the base layers have been frozen. We trained each neural network on 30 epochs and categorical accuracy has been used as a measure for validation. Fig. 3 depicts the architecture of the proposed model.

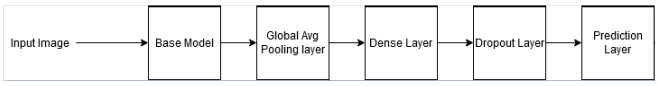

Fig. 3. Model architecture

III. EVALUATION METRICS

Several evaluation metrics could be used for a classification task but here we take into consideration accuracy, precision, and recall. The recall is the most important measure here. TP and TN represent True Positives and True negatives. Similarly, False Positives and False Negatives are represented by FP and FN respectively.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = (TP) / (TP + FP) \quad (2)$$

$$\text{Recall} = (TP) / (TP + FN) \quad (3)$$

IV. RESULTS

To avoid overfitting and handle class imbalance, techniques like dropout, data augmentation, class-weights have been used. 4 different architectures have been tried in the place of the base model to find the final best model. The results can be summarized in Table 1. ResNet50 has been selected as the final model after comparing it with other models as shown in Table 1 based on the evaluation metrics. For all models, learning rate reduces on the plateau with initial learning rate 0.01, Adam optimizer and we saved the best parameters for those models using the checkpoint. After

an evaluation of the test set, our final model ResNet50 has achieved an accuracy of 90%, weighted average precision of 0.89, and recall of 0.90. Weighted average precision and weighted average recall have been used as they consider class frequency rather than giving equal weights to all classes. Table 1 represents the accuracies of various architectures.

TABLE I. TESTED ARCHITECTURES

Architecture	Accuracy	Weighted average precision	Weighted average recall
ResNet50	90	89	90
MobileNet	87	86	87
Xception	84	84	84
InceptionV3	85	85	85

While selecting a final model, our goal was to select a model that produces minimal false negatives as it is a crucial factor in the medical domain. To minimize false negatives, we need to maximize recall. Further, we summarized the model performance concerning each class based on its recall as shown in Table 2. AKIEC is a potentially malignant class while BCC and MEL are malignant. Despite both ResNet50 and Xception have a similar individual class recall for the malignant classes, ResNet50 has a better weighted average recall of 0.9 which is better than Xception at 0.84.

TABLE II. TESTED ARCHITECTURE

	akiec	bcc	bkl	df	mel	nv	vasc	WAR
1	0.42	0.6	0.68	0.17	0.41	0.99	0.82	0.9
2	0.19	0.57	0.59	0	0.36	0.97	0.91	0.87
3	0.46	0.6	0.28	0.33	0.41	0.94	0.64	0.84
4	0.27	0.67	0.4	0.17	0.51	0.95	0.73	0.85

In Table II: 1,2,3,4 represent Resnet50, MobileNet. Xception and InceptionV3 respectively and WAR denote Weighted average recall.

In comparison with similar works as shown in Table 3, we found that Esteve et al [14] achieved an accuracy of 72.1 percent, Mohammed et al [15] achieved an accuracy of 92.70 percent using MobileNet, Gupta et al [16] used EfficientNetB1 and achieved 94.00 percent accuracy. Muresan et al [17] used an ensemble of Inception-ResNet and reported an accuracy of 83.96 percent while Nugroho et al [18] worked on a model from scratch, achieved 78.00 percent accuracy. One of the early works of using deep learning for melanoma classification was proposed by Esteve et al [14]. The efficiency of their model was tested by 21 board-certified dermatologists. Our model was only behind Mohammed et al [15] among other works, not only in terms of accuracy but also weighted average precision and recall. The comparative analysis of accuracies of proposed model and existing models presented in table 3.

TABLE III. COMPARISON WITH OTHER WORKS

Work	Architecture	Accuracy	Precision	Recall
[15]	MobileNet	92.70	87.00	81.00
[16]	EfficientNet B1	94.00	94.00	94.00
[17]	Inception-ResNet	83.96	72.00	69.29
[18]	From Scratch	78.00		
Our Model	ResNet50	90.00	89.00	90.00

V. CONCLUSION

In this paper, we have used transfer learning as a technique to train neural networks to classify seven types of skin cancer. Even though skin cancer can be easily cured when detected at an early stage, detection has been a challenging task. We propose a model with ResNet50, MobileNet, Xception, and InceptionV3 as the base models for the classification of skin lesion images in the HAM10000 dataset.

We tried to tackle the class imbalance by implementing additional techniques such as data augmentation and class weights while training. After examining them based on our evaluation metrics, ResNet50 was selected as the final model. Our final Model based on ResNet50 has achieved a multi-class accuracy of 90 percent with a weighted average precision of 0.89 and a weighted average recall of 0.90. MobileNet can be a trade-off between performance and size when there is a need to deploy on mobile or web applications. Also, the output of the base layers can be used as the input to traditional machine learning algorithms like SVM and other ensemble models rather than simple dense layers.

The use of such machine learning models as the top layers can help us with an interpretation of the model. Learning to understand how CNNs interpret images can be useful for creating a better model for the problem we have at hand. Neural networks are usually considered black-boxes and we cannot derive a particular reason for how they have performed. Interpreting them can be helpful for further studies, especially in the medical domain. Performing segmentation on the skin lesion images and using them as an input to the classification model might be a way to reduce noise in the input data. This method helps in increasing the performance of the model. With larger computational and memory resources, the study can be extended by examining larger architectures and performing deeper fine-tuning.

REFERENCES

[1] https://www.cancer.gov/about-cancer/understanding/statistics (Date accessed: 09-Nov-2020)

[2] https://www.cancer.gov/types/skin/hp/skin-prevention-pdq (Date accessed: 09-Nov-2020)

[3] https://www.who.int/news-room/q-a-detail/radiation-ultraviolet-(uv)-radiation-and-skin-cancer (Date accessed: 09-Nov-2020)

[4] https://www.cancer.gov/types/skin/research (Date accessed: 09-Nov-2020)

[5] https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/isicArchive (Date accessed: 09-Nov-2020)

[6] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," Lancet Oncol., vol. 3, no. 3, pp. 159–165, Mar. 2002.

[7] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, "Rethinking the Inception Architecture for Computer Vision". arXiv:1512.00567

[9] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

[10] François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions" arXiv:1610.02357

[11] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, "Big Transfer (BiT): General Visual Representation Learning" arXiv:1912.11370

[12] Philipp Tschandl, Cliff Rosendahl, Harald Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions" arXiv:1803.10417

[13] Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson, "How transferable are features in deep neural networks?". arXiv:1411.1792

[14] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[15] E. H. Mohamed and W. H. El-Behaidy, Enhanced Skin Lesions Classification Using Deep Convolutional Networks, Proc. - 2019 IEEE 9th Int. Conf. Intell. Comput. Inf. Syst. ICICIS 2019, pp. 180188, 2019, doi: 10.1109/ICICIS46948.2019.9014823.

[16] H. Gupta, H. Bhatia, D. Giri, R. Saxena, and R. Singh, Comparison and Analysis of Skin Lesion on Pretrained ArchitecturesComparison and Analysis of Skin Lesion on Pretrained Architectures no. July 2020, doi: 10.13140/RG.2.2.32161.43367.

[17] H. B. Muresan, Skin Lesion Diagnosis Using Deep Learning, Proc. - 2019 IEEE 15th Int. Conf. Intell. Comput. Commun. Process. ICCP 2019, pp. 499506, 2019, doi: 10.1109/ICCP48234.2019.8959661

[18] A. A. Nugroho, I. Slamet, and Sugiyanto, Skins cancer identification system of HAMl0000 skin cancer dataset using convolutional neural network, AIP Conf. Proc., vol. 2202, no. December, pp. 06, 2019, doi: 10.1063/1.5141652.