New Compact Deep Learning Model for Skin Cancer Recognition

Phillip Ly
Department of Computer Science
California State University
Fullerton, California 92831
phillipdly@csu.fullerton.edu

Doina Bein
Department of Computer Science
California State University
Fullerton, California 92831
dbein@fullerton.edu

Abhishek Verma
Dept. of Computer Science
New Jersey City University
Jersey City, NJ 07035
averma@njcu.edu

***Abstract*— Deep learning neural networks have made significant progress in image analysis and have been used for skin cancer recognition. Early detection and proper treatments for malignant skin cancer cases are vital to ensure high survival rate in patients. We present a novel deep learning based convolutional neural network (CNN) model for generating compatible models on mobile platforms such as Android and iOS. The proposed model was tested on the grand challenge PHDB melanoma dataset. The best performing proposed model excels in the following ways: (1) it outperforms the baseline model in terms of accuracy by 1%; (2) it consists of 60% fewer parameters compared to the base model and thereby it is more efficient on mobile platforms. Furthermore, the model is more compact and retains high accuracy without the need to be downsized; (3) in conjunction with advanced regularization techniques such as dropout and data augmentation, the proposed CNN model excelled when implemented on state-of-the-art frameworks such as Keras and TensorFlow. Additionally, we were able to successfully deploy it on the iOS and Android mobile systems. The proposed model could also be lucrative towards other datasets for image classification on mobile platform.**

Keywords—deep learning; skin cancer; melanoma; neural network; CNN; PHDB; mobile systems

I. INTRODUCTION

Skin cancer is the most common type of cancer in the United States [1]. According to the American Cancer Society, there were about 5.4 million new cases of skin cancer in 2017. In fact, the number of new cases for skin cancer exceeds the combined total of new cases for prostate cancer, breast cancer, lung cancer, and colorectal cancer [2]. Malignant melanoma is a prevalent type of cancer that is especially deadly. Therefore, the goal of this research is to propose a new convolutional neural network based deep learning model that can detect melanoma in its early stage and can be used on mobile platforms.

Early detection and proper treatments for new malignant skin cancer cases are very important to ensure high survival rate. For instance, the survival rate for melanoma decreases from 99% to 14% in more advanced stages. According to IMS Health, there are only 9,600 dermatologists and 7,800 dermatology practices to serve 323 million people in the U.S [3]. Therefore, it is imperative to extend the reach of such essential diagnostic care by orchestrating deep learning for skin cancer classification.

Professional dermatologists have established the ABCDEs

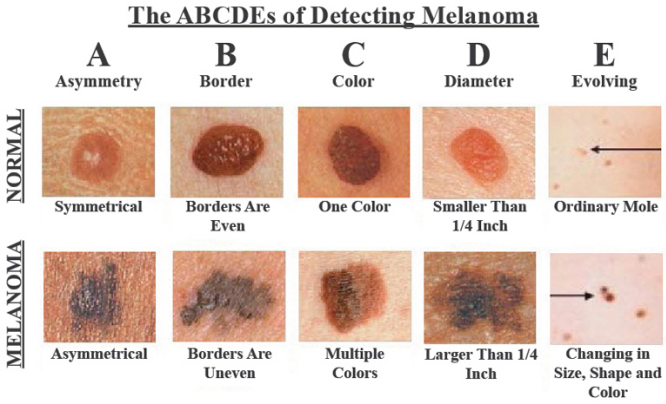

Figure 1. The popular ABCDEs for classifying melanoma.

(Asymmetrical shape, Border irregularities, Color, Diameter, and Evolution) in Fig. 1 as the standardized descriptions to assist with visualizing common features of malignant melanoma cases. One of the main challenges of classifying malignant skin lesions is due to the sheer amount of variations across the diverse skin tones from people of different ethnic backgrounds [4]. Moreover, according to Dr. Darrick Antell, some melanoma cases can have pinkish, white, red, or even clear appearances rather than the usual dark pigment melanin [5]. In recent years, new breakthroughs in the development of convolutional neural networks (CNNs) have allowed computers to outperform dermatologists in skin cancer classification tasks [6]. The next major step is to further improve the accuracy of melanoma detection.

This paper is organized as follows. Section II presents background discussion. Section III provides the description of dataset used in our experiments. Section IV includes details of our proposed deep learning model. The experimental approach and results are introduced in Section V. Finally, the conclusion and future work are discussed in Section VI.

II. BACKGROUND AND RELATED WORK

Researchers have developed a myriad of skin cancer classification systems by employing large CNN architectures such as InceptionV3 [6], Microsoft ResNet-152 [7], VGG16 [4], GoogleNet, and VGG-19 [8]. These large CNN architectures are designed to have deeper layers as parts of more complex neural networks in order to maximize top-1 and

Table 1. Comparison of Results from Various Deep Learning Researches for Skin Cancer Classification

Author	Best Classifier (Accuracy)	Classification Technique	Neural Network	Input Size (Pixels)	Dataset Type	Size of Model (MB)	Mobile Platform
Authors of Proposed Model	86%	Train from scratch	Model_A	224 x 224	Balanced	29	iOS and Android
Esteva [6]	72.1%	Transfer learning	Inception V3	299 x 299	Imbalanced	Not reported	Not reported
Kalouche [4]	78%	Transfer learning	VGG - 16	256 x 256	Imbalanced	Not reported	Not reported
Han [7]	57.3%	Transfer learning	ResNet - 152	224 x 224	Imbalanced	Not reported	Not specified

top-5 accuracies [9][10][11]. However, large-scale CNN architectures can consist of hundreds of millions of parameters which can lead to overfitting when training on small and imbalanced datasets [12] such as skin cancer datasets. Large CNN architectures also generate sizable models (larger than 100 MB) and impede deployment to resource-constrained mobile systems. Typically, the size of deep learning models for malignant melanoma recognition should be as small as possible (and retain high accuracy) to ensure that the models can be deployed on iOS or Android systems. Furthermore, the input size of images must be resized to 224 x 224 pixels for the deep learning models to work properly on mobile systems.

Table 1 shows the comparisons between results of proposed model and results from [4][6][7]. An interesting aspect regarding the listed comparisons is that we have created a new CNN model (with 86% top-1 accuracy) and trained on a balanced dataset from scratch, whereas other researchers have selected to use transfer learning techniques via large-scale CNNs on imbalanced datasets. Furthermore, based on the published papers, other researchers have not reported the size of their models and it is unknown if they deploy their models on mobile platform such as iOS. The proposed deep learning model is compact and takes 29 megabytes (MB) in size.

Large CNNs are pretrained on the ImageNet dataset which contains 1.4 million images of common objects and thousands of different classes. Traditionally, transfer learning techniques are used to solve image classification problems such as classifying cats and dogs because the ImageNet dataset contains animal classes that include several types of cats and dogs. Contrarily, the ImageNet dataset comprised of images that are very different in comparison to skin lesion images. Thus, it is a safer approach to train deep learning models (for malignant melanoma recognition) from scratch because it allows us to have more control over the exact inputs and outputs of CNNs. Such an important but subtle aspect of training CNNs from scratch allows us to have a more direct impact on the ways CNNs learn about the features from the input of data and how they generate predictions for classification. For mobile applications to successfully perform early detection of malignant melanoma, we need to maximize the accuracy of the deep learning model as much as possible.

Another important aspect regarding Table 1 is that other researchers used imbalanced datasets. Typically, CNN models that are trained using imbalanced datasets are likely to be biased toward the classes that consist of larger number of images [4]. In other words, deep learning models for skin cancer classification are forced to make the wrong predictions in favor of benign cases because the CNNs learn more about the features of benign images during training. Thus, we must balance the numbers for benign and malignant cases during construction of skin lesion datasets. As we obtain more malignant images, we can then add an equal number of benign images to our dataset. Most skin lesion datasets contain higher numbers of benign images compared to malignant images. Thus, it is intuitive to use a balanced dataset to acquire the model with the best possible accuracy.

We use Keras [15] to train our model. Note that in Keras version 2.0, the Keras team has removed metrics such as recall,

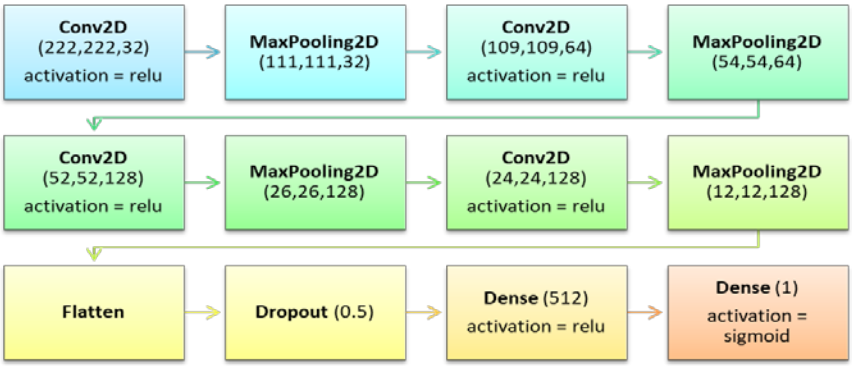

Figure 2. Visualization of baseline CNN model (Model_B).

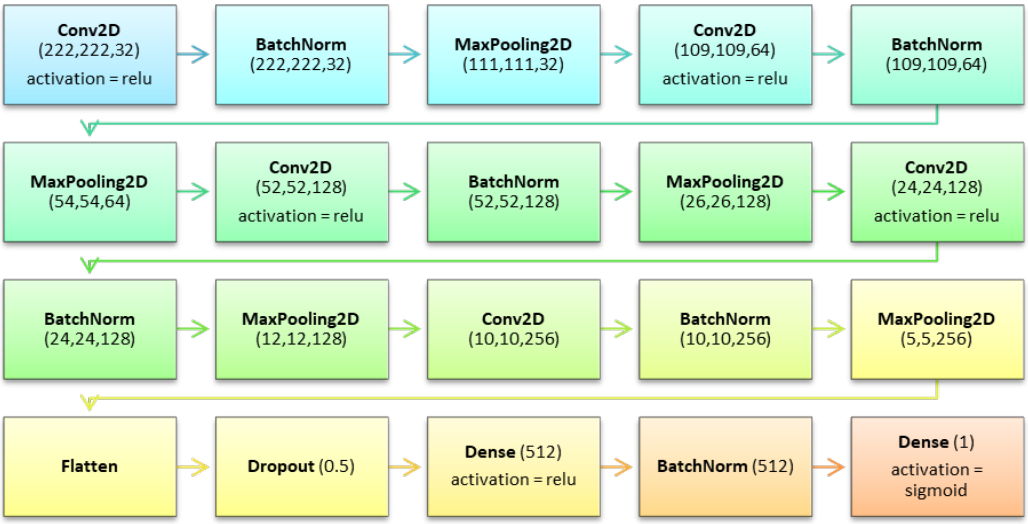

Figure 3. Architecture of the best proposed CNN model (Model_A).

precision, and fmeasure to implicitly promote the use of accuracy as the main metric for CNNs that are trained on balanced datasets [17]. Ultimately, large deep learning models require a downsizing methodology that would decrease the overall performance of proposed skin cancer classification on mobile platform. Therefore, we propose a more efficient CNN architecture that can accommodate deep learning on mobile systems.

III. Description of our Dataset

The proposed methodology in Section IV entails the construction of a novel convolutional neural network that is trained on a composite dataset of skin lesion images. This composite dataset comprises of several publicly available datasets, which include ISIC Archive [19], Dermnet NZ [20], MED-NODE [21], and PH2 [22]. A composite skin cancer dataset titled "PHDB" is created from the combination of listed datasets.

The dataset is divided into two parts representing the benign and malignant classes. Each part contains a balanced number of malignant or benign skin lesion images. In other words, 50% of the PHDB dataset is comprised of benign images and the remaining 50% are malignant images. The PHDB dataset is partitioned as follows: 80% for training, 10% for validation, and 10% for testing. We have manually inspected all images after they are partitioned to ensure that the various versions of images generated via data augmentation of original skin lesion image do not overlap across the train/validation/test sets. Furthermore, we have ensured that the skin lesion images in the test and validation sets are high-quality and biopsy-proven.

At run time, the PHDB dataset is increased to about 80,192 images for training via a combination of various regularization techniques. Regarding data augmentation technique from related work, Esteva [6] performed random image rotations between 0° to 359°. Likewise, Han [7] performed similar image augmentation techniques to increase his dataset (by 20 to 40 times the original total of 20,826 images). All skin lesion images from the specified datasets are publicly available and are free to use for educational and research purposes.

IV. Proposed Deep Learning CNN Model

In Fig. 2 we present the baseline CNN. The baseline CNN architecture has a sequence of alternating Conv2D and MaxPooling2D layers that form the core building blocks of modern CNNs. The first convolutional layer takes in 224 x 224 skin lesion images. The last dense layer of the baseline CNN contains a single unit with sigmoid activation in order to output the resulting classes (benign and malignant).

Fig. 3 shows the architecture of proposed CNN (Model_A). We have added additional batch normalization layers after every Conv2D layers and dense layers to facilitate normalization of distribution for each batch of data. The proposed CNN has five sets of Conv2D, BatchNorm, and MaxPooling2D. Based on many experiments, these sets represent the most optimal CNN setup to process skin lesion images from the PHDB dataset and attains the highest possible accuracy using the train from scratch method. Furthermore, Fig. 3 shows the proposed CNN that attains higher accuracy and has 5,860,736 fewer parameters in comparison to the baseline model. More specifically, the baseline CNN has a total of 9,679,041 parameters, whereas the aggregate number of parameters of the proposed model is 3,818,305.

The PHDB dataset consists of high resolution images that necessitate a larger neural network with more layers and capacity. As seen in Fig. 2, the Conv2D and MaxPooling2D layers help to augment the CNN's capacity. This baseline CNN has attained an accuracy of 85%. Furthermore, the size of the feature maps is reduced such that they would be compatible as inputs into the Flatten layer. For example, when skin cancer images are fed into the CNN, the initial size of the feature maps in the first Conv2D layer is 224 x 224. Then, as the CNN processes data through the alternating Conv2D and MaxPooling2D layers, the feature maps are transformed into

Table 2. Comparison of Results from Various Proposed Models and the Baseline Model (Model_B).

Proposed Models	Accuracy (Trained for 50 Epochs)	Structure of Feature Maps	Total Parameters	Dropout	Activation	Max Pooling	Batch Norm	Optimizer	
		Conv2D, MaxPool, BatchNorm	Trainable	Rate and Location	Type	Total Layers	Total Layers	Type	Learning Rate
Model_A	86%	32x3 => 64x3 => 128x6 => 256x3	3,818,305	0.5 Before	ReLU	5	5	RMSprop	0.0001
Model_B (Baseline)	85%	32x2 => 64x2 => 128x4	9,679,041	0.5 Before	ReLU	4	0	RMSprop	0.0001
Model_C	84.5%	32x3 => 64x3 => 128x6 => 256x3	3,818,305	0.5 Before	ReLU	5	5	Adam	0.0010
Model_D	85.5%	32x3 => 64x3 => 128x6 => 256x3	3,818,305	0.5 After	ReLU	5	5	RMSprop	0.0001

the size of 12 x 12 in the last MaxPooling2D layer right before the Flatten layer.

An important aspect regarding the proposed CNN is that the depth of the feature maps gradually becomes larger (from 32 to 128). In contrast, the size of the feature maps decreases from 224 x 224 to 12 x 12 as data traverses from input layer (first Conv2D layer) to output layer, Dense (1). Moreover, the proposed CNNs in this section are used to solve binary-classification problem (malignant or benign). Thus, the Dense (1) in Fig. 3 contains a single unit with sigmoid activation. This unit is responsible for encoding the probability, which CNN utilizes to determine either malignant or benign category of the input images.

As shown in Fig. 3, the proposed CNN utilizes batch normalization layers and consists of 3,818,305 parameters in total. The proposed CNN is significantly smaller than the baseline CNN. Furthermore, the modified CNN has an additional Conv2D layer with the output shape of (10, 10, 256). This helps to further simplify the representations and reduces the overall number of parameters. Most importantly, the accuracy of the modified CNN (Model_A) increased to 86%

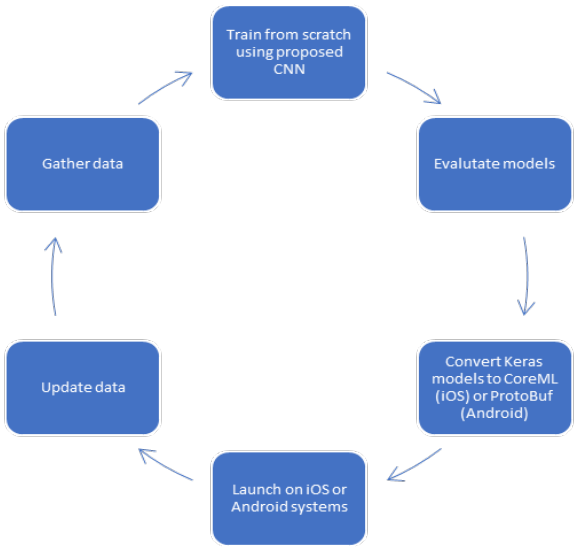

Figure 4. General workflow for proposed models.

using the same training and data augmentation configurations as the baseline CNN (Model_B). Furthermore, the depth of the feature maps (the last number of the feature map's shape in layers as shown in Fig. 3) gradually increases (from 32 to 256) in a way that is like larger CNN architectures such as VGG-16 [10].

Table 2 shows the comparison between various proposed models and the baseline model in terms of feature maps structure, hyperparameters, and parameters. All models were trained for 50 epochs. The structure of feature maps column represents the shape of the feature map at each layer of proposed CNN as shown in Fig 3. The dropout layer (0.5) was placed before or after the Flatten layer as a part of experiments.

The proposed CNN architecture is very versatile in terms of scalability. It can be used to solve other binary classification problems (such as classifying cats and dogs). We can add or remove additional sets of Conv2D, BatchNorm, and MaxPooling layers as needed (and tweak the size of the feature maps) to solve many classification problems that use combination of various datasets.

The batch normalization layers are a special type of layer that was first introduced by Ioffe and Szegedy [13]. Batch normalization layers perform normalization of data in an adaptive manner by taking into consideration the alterations of mean and variance during training. Internally, it functions by preserving the moving average of the batch-wise mean and variable of the processed data during training. Moreover, batch normalization layers are pivotal for building deep neural networks because they help to facilitate the gradient propagation in a similar way to residual connections [12]. In the case of building CNNs for skin cancer classification, batch normalization layers help to speed up the training process [14].

Figure 4 displays the general workflow for the proposed CNN. The CNN models generated through Keras are converted to Core ML models and deployed on iOS mobile systems. Likewise, Keras models can be transformed into protocol buffer files and deployed on Android systems.

V. Experimental Environment and Results

All experiments are performed on an Ubuntu 16.04 server equipped with seven NVIDIA GeForce GTX 1080 Ti (12GB of VRAM each). Moreover, this server also includes two Intel

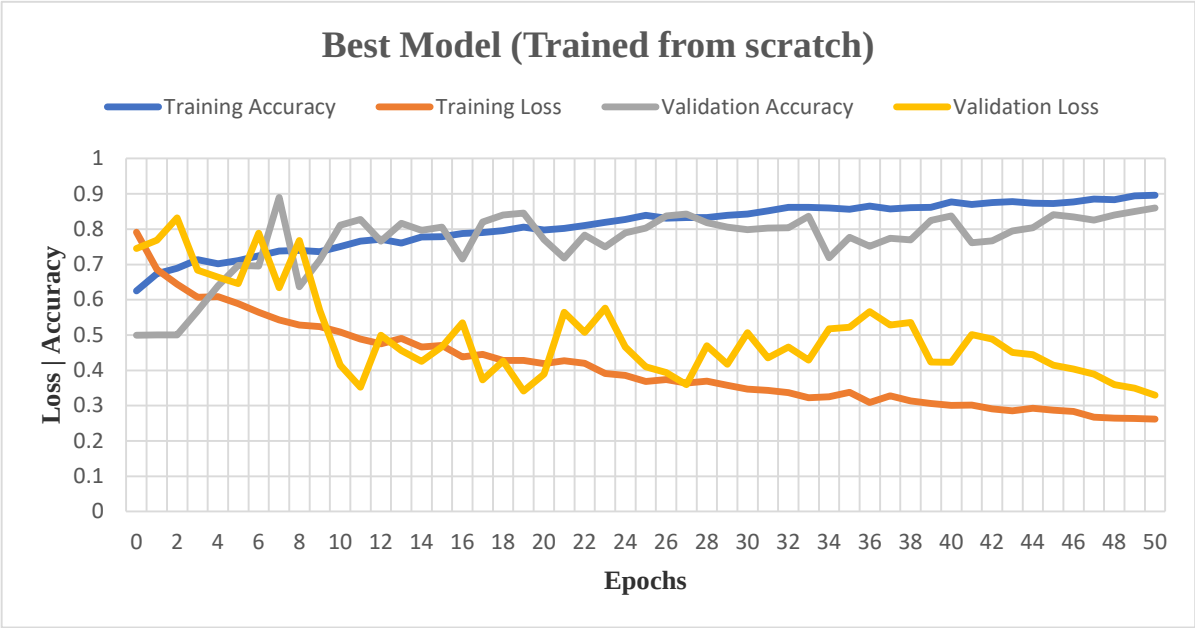

Figure 5. Training and validation performance of the best proposed deep learning model (Model_A).

Xeon processors E5-2630 v4 2.20GHz which have 2 threads per core, 10 cores per socket, and 256 GB of main memory.

A. *Configuration for Training*

The configuration to train the model is as follows: RMSProp is set as optimizer, 0.0001 as learning rate, loss function is binary_crossentropy, and the metrics is set to 'acc'. The previously mentioned hyperparameters are configured via the compile method in Keras [15]. Google's TensorFlow [16] machine learning library is used alongside Keras as the backend.

B. *Configuration for Data Augmentation*

The configuration for data augmentation is as follow: rotation_range = 45, width_shift_range = 0.05, height_shift_range = 0.05, shear_range = 0.1, zoom_range = 0.1, horizontal_flip = True, and fill_mode = 'nearest'.

Finally, the batch generator is instantiated for the declarations of hyperparameters that would determine the training duration of the proposed models. The configurations are assigned as follow: steps_per_epoch = 100, epochs = 50, and validation_steps = 50.

C. *Analysis of Results*

Figure 5 displays the experimental results of best performing proposed model (Model_A) with 86% top-1 accuracy. The duration of training lasted for 50 epochs; models seemed to converge around that mark. Regularization techniques such as dropout (0.5) and data augmentation techniques were heavily relied upon to combat the overfitting problem. The process to build the PHDB dataset was initiated back in June 2017. Thus, the valuable HAM10000 [23] dataset was not publicly available at the time. Furthermore, we are working with a balanced dataset as input and proposed models are trained from scratch. Thus, the validation accuracy curve is not expected to be perfectly smooth compared with results that

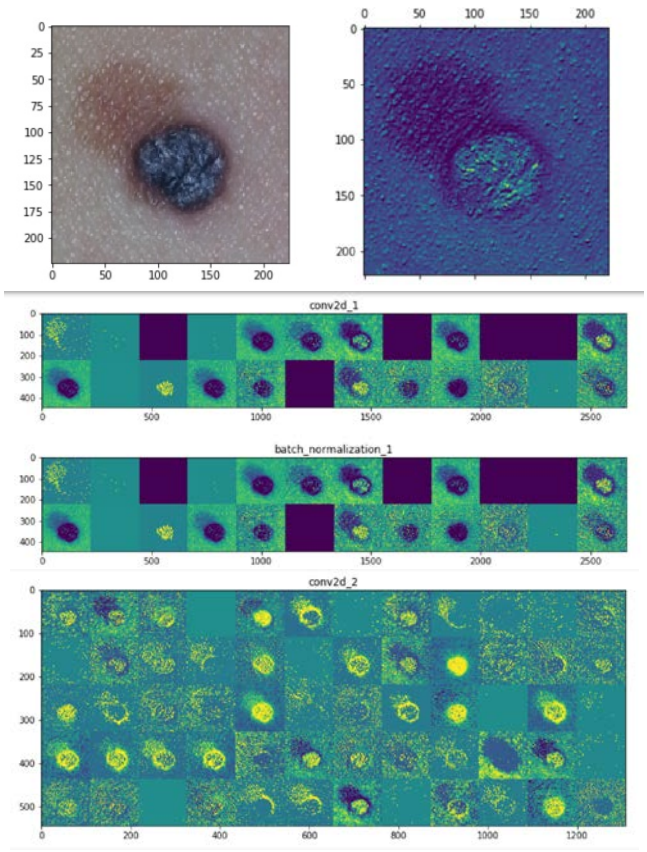

Figure 6. Visualization of feature maps generated from proposed CNN model with three dimensional (width, height, depth) channels for a skin lesion image from PHDB dataset.

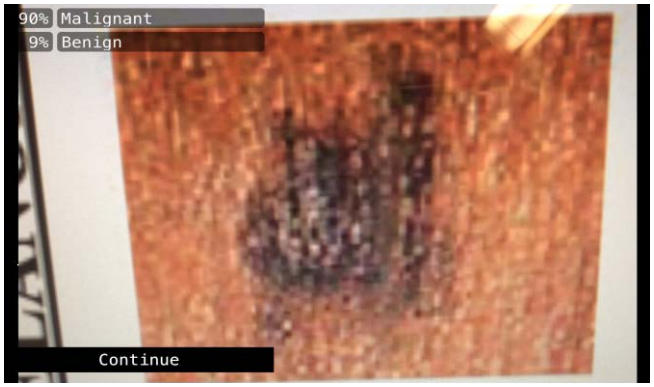

Figure 7. Demonstration of malignant melanoma recognition on an iOS mobile system.

are generated from transfer learning techniques. However, in future, with the availability of additional skin lesion images from the HAM10000 and Dermofit [24] to the PHDB dataset, we expect the validation accuracy curve to further smooth out. Moreover, the training accuracy and the training loss curves from Fig. 5 for the best performing model are very smooth.

The hyperparameters to perform data augmentation are adjusted as follow: rotation_range = 45, width_shift_range = 0.03, height_shift_range = 0.03, shear_range = 0, zoom_range = 0.1, horizontal_flip = True, and fill_mode = 'nearest'. The learning rate remains at 0.0001 with RMSprop optimizer.

Fig. 6 shows the visualization of feature maps of a skin lesion image as it traverses through the proposed CNN layers. This visualization assists us to understand how the layers of proposed CNN model transform the input of skin lesion images and generate the corresponding predictions.

Figure 7 displays an actual demonstration of the deep learning model, which represents a much more reliable way to validate the performance of the proposed CNN model. Additionally, we can successfully deploy the model on the iOS and Android mobile systems. Fig. 7 shows a demonstration of malignant melanoma recognition on an iPad with iOS version 11.4 installed. The glare from the fluorescent lamp is introduced on the top right corner of the image in order to test the robustness of the deep learning model. Malignant melanoma classifier managed to correctly classify the image as malignant with 90% confidence.

We strongly believe in deploying deep learning models on mobile systems and testing them in the real world represents a much more reliable methodology than any generic metrics when it comes to measuring how a model performs a certain task. The rationale is that future users will use the products in real-world situations. The malignant image in Fig. 7 is a part of Fig. 1. We have correctly, in the real world, classified 7 out of 8 images from columns A through D. The images from Fig. 1 are new images that the proposed CNN has never seen before. Moreover, the images from column E are ignored due to the insufficiently available features for the classifier to correctly analyze and generate reliable prediction.

D. Validation Accuracy Graph

Figure 8 shows the comparison of validation accuracy for the models listed in Table 2. Model_A has the highest

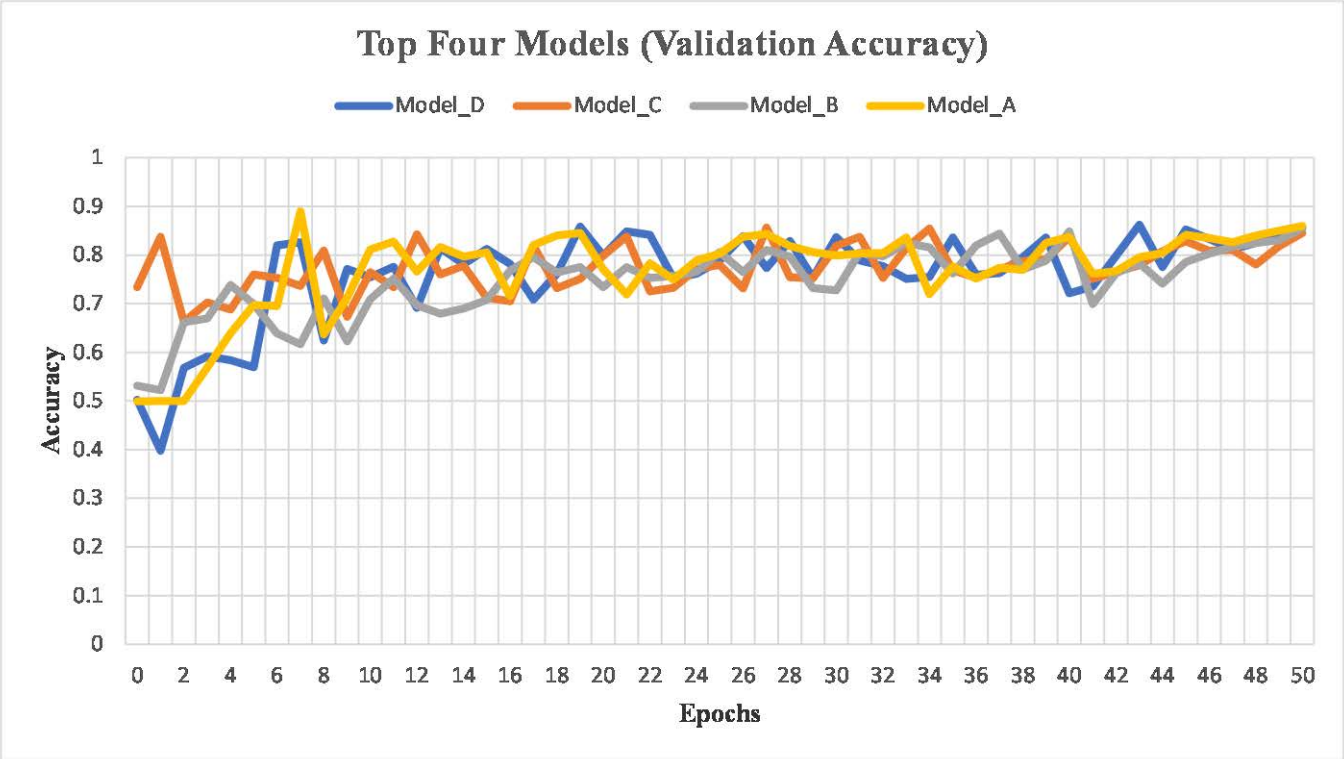

Figure 8. Validation accuracy graph for best four proposed models.

Actual	**Malignant**	**120**	16
	Benign	22	**114**
		Malignant	**Benign**
		Predicted	

Figure 9. Confusion matrix for best model (Model_A).

accuracy at 86%. It was trained for approximately 43 hours. All the other models were trained between 30 to 45 hours depending on the distributed workload on the Ubuntu server. Furthermore, we could have selected the smoother training accuracy curves to display in Fig. 8. However, the validation accuracy is used instead because it represents a more reliable metric than the training accuracy to measure the performance of deep learning models. Many experiments were carried out and those four represent the best performing models.

E. Confusion Matrix

Figure 9 presents the confusion matrix for Model_A. The *sensitivity*, i.e., correctly classified instances of malignant lesions is 88.23% whereas the *specificity*, i.e., correctly classified instances of benign lesions is 83.82%.

VI. Conclusion and Future Work

This paper presents a novel convolutional neural network based deep learning model for generating mobile compatible models. The proposed model outperforms the baseline model in terms of accuracy. We trained the network from scratch. Furthermore, it has the highest accuracy for malignant melanoma recognition in comparison to previous researches. The new model consists of less number of parameters, making it more efficient on mobile platforms, thereby making it more compact and highly accurate without the need to be downsized. In conjunction with advanced regularization techniques such as dropout and data augmentation, the proposed CNN model has excellent performance when implemented on state-of-the-art frameworks such as Keras and TensorFlow.

A critical area of future work is to incorporate new data to acquire models with even higher accuracy. New datasets such as HAM10000 and Dermofit would be added to the PHDB dataset in the future. Furthermore, new generation of neural networks inspired by Google's MobileNetV2 and Geoffrey Hinton's Capsule Networks may offer further possibilities of improving the accuracy [18]. Ultimately, we expect deep learning and mobile systems to be a promising combination that could transform healthcare for billions of people.

Acknowledgment

Phillip Ly acknowledges the support by the Associated Students, CSUF, Inc. (ASI) for the ASI Student Research Grant under award number 18-010.

References

[1] G. Guy, et. al., "Vital signs: Melanoma incidence and mortality trends and projections—United States, 1982–2030," MMWR Morb Mortal Wkly Rep, vol. 64, no. 21, pp. 591-596, 2015.

[2] American Cancer Society, "Don't Fry: Preventing Skin Cancer," [Online]. Available: https://www.cancer.org/research/infographics-gallery/skin-cancer-prevention.html. [Accessed 16 April 2018].

[3] M.P. Seraly, "A new patient management service for dermatologists," [Online]. Available: http://iagnosis.com/sites/default/files/DermatologistOnCall-Whitepaper.pdf. [Accessed 13 April 2018].

[4] S. Kalouche, "Vision-Based Classification of Skin Cancer using Deep Learning," 2016. [Online]. Available: https://www.semanticscholar.org/paper/Vision-Based-Classification-of-Skin-Cancer-using-Kalouche/b57ba909756462d812dc20fca157b3972bc1f533.

[5] A. Venosa and D. Antell, "More Than Moles: When Melanoma Doesn't Look Like You Think It Should," 2018. [Online]. Available: https://blog.skincancer.org/2017/10/04/more-than-moles-melanoma/. [Accessed 17 April 2018].

[6] A. Esteva, et al., "Dermatologist - level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115-118, 2017.

[7] S. Han, et. al., "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, Vol. 138 (7), pp. 1529-1538, 2018.

[8] H. Liao, "A Deep Learning Approach to Universal Skin Disease Classification," 2016. [Online]. Available: https://pdfs.semanticscholar.org/af34/fc0aebff011b56ede8f46ca0787cfb1324ac.pdf. [Accessed 11 April 2018].

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385. [Accessed 16 April 2018].

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556. [Accessed 15 April 2018].

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015. [Online]. Available: https://arxiv.org/abs/1512.00567.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016. Available: https://arxiv.org/abs/1610.02357

[13] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *Proc. of the 32nd Int. Conf. on Machine Learning*, 2015.

[14] A. Geron, "Hands-On Machine Learning with Scikit-Learn & TensorFlow," O'Reilly Media, Inc., 2017.

[15] Keras. 2018. [Online]. Available: http://keras.io. [Accessed 16 April 2018].

[16] TensorFlow. 2018. [Online]. Available: https://www.tensorflow.org/. [Accessed 16 April 2018].

[17] Keras Team. 2018. [Online]. Available: https://github.com/keras-team/keras/wiki/Keras-2.0-release-notes. [Accessed 17 August 2018].

[18] Hinton, G., "Dynamic Routing Between Capsules" 2017. [Online]. Avalable: https://arxiv.org/abs/1710.09829. [Accessed November 2017]

[19] ISIC Archive. [Online]. Available: https://www.isic-archive.com/. [Accessed 16 June 2017].

[20] Dermnet NZ. 2018. [Online]. Available: https://www.dermnetnz.org/. [Accessed 16 June 2017].

[21] I. Giotis, et. al., "MED-NODE: A computer assisted melanoma diagnosis system using non-dermoscopic images," *Expert Systems with Applications*, vol. 42 (19), 2015.

[22] PH2. [Online]. Available: https://www.fc.up.pt/addi/ph2%20database.html. [Accessed 16 June 2017].

[23] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions," 2018. Available: https://arxiv.org/abs/1803.10417

[24] L. Ballerini, et. al., "A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions," *Color Medical Image Analysis, Lecture Notes in Computational Vision and Biomechanics*, 2013.