Skin Cancer Detection Using Convolutional Neural Networks*

Mattia Galanti[1], *Gilliean Lee*[2], *Joshua John*[2]
[1] *School of Computing*
Clemson University
Clemson, SC 29634
{mattia.galanti}@clemson.edu
[2] *Department of Mathematics & Computing*
Lander University
Greenwood, SC 29649
{glee,joshua.john}@lander.edu

Abstract

This paper focuses on the development of classifiers capable of detecting a skin cancer(s) given dermoscopic images. The dataset used for the training is a part of the 2019 ISIC Challenge, and consists of more than 25,000 labeled dermoscopic images. Specifically, classifying dermoscopic images accounts for nine different diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma, some of which are benign. We have developed classifiers –a binary classifier and a multiclass classifier –on the Google Cloud Platform using Convolutional Neural Networks (CNNs). To prevent the classifiers from overfitting and to achieve higher accuracy even with the smaller training data size, we use image data augmentation. The binary classifier achieved an accuracy of 79% with 220 epochs of training, and the multiclass classifier' s accuracy is 72% with 200 epochs.

1 Introduction

Skin cancer, more than any other type of cancer, causes more death than heart disease in wealthy countries [3] and is increasingly the most common form of cancer in the United States [14]. In 2019, it is estimated that 7,230 deaths are attributed to melanoma alone. However, skin cancer is also one of the most treatable types of cancer. The five-year survival rate for melanoma patients is 99% if it is detected and treated before it spreads to the lymph nodes; thus, early detection is crucial. The warning signs of skin cancer include changes in size, shape, and color of moles or skin lesions [14]. Despite all this, it is very unlikely that during a routine check-up a skin lesion will be classified as a possible skin cancer. A fair number of patients diagnosed with melanoma had visited their regular doctors in the two years or so preceding the diagnosis [21]. In most cases, the melanoma was not diagnosed, and some patients were incorrectly cleared after having reported concern about skin lesions [21]. Over the last decade, the growing focus has been put on identifying screening obstacles in order to implement more tailored strategies that promote early detection efforts [21]. Locating skin cancer is not generally a part of the main practice for essential care. The chance of detecting a clinically imperative early skin cancer is low, thus practitioners are often not able to detect skin lesions at their earliest stages of development. Yet physicians never face the consequences of missing an early skin cancer diagnosis. Further, there are not many specialists who have the expertise to suitably diagnose potentially cancerous lesions [21].

Skin cancer detection is being revolutionized by deep learning. Deep learning is a subset of machine learning and a specific form of artificial neural network (ANN), similar to the multilayered human cognition system [12]. Deep learning methods have changed the investigation of natural images and videos, and similar examples are emerging with the investigation of biomedical data. Deep learning has been used to classify lesions and nodules, and to localize organs, regions, and landmarks [4]. In numerous areas of medicine, deep learning technology is used widely. One area of its application includes breast cancer detection. Here, traditional detection methods are often expensive, painful, and inaccurate with respect to measuring the location and size of the tumor. To overcome these deficiencies, neural network models for early breast cancer detection have been used [1]. Additionally, research on leukemia classification in later years has been primarily based on computer vision techniques. Machine learning techniques, such as K-means clustering, have been implemented in order to detect and classify blood cells in images. There are two advantages to using Convolutional Neural networks (CNNs): not only do they diminish the processing time by allowing us to skip most of the pre-processing steps, but they also are capable of extrapolating features that are more robust than the customary statistical features [19]. In recent years, deep learning has been

revolutionizing liver lesion segmentation, which is a fundamental step in the diagnosis of liver cancer. While manual segmentation is time-consuming and prone to error, scholars have designed a two-step U-Net approach in order to automatize such segmentation, and they have reported a very high accuracy rate (Dice score above 0.94) on their test data [7]. The Dice score, or Sørensen–Dice coefficient, is a common metrics for the evaluation of segmentation tasks in medical imaging [2]. All this is only a portion of the vast progress made in recent years in the biomedical field with regard to the detection of tumors.

Our research aims to develop a binary skin cancer classifier and a multi-class classifier using deep learning technology by training them with dermoscopic skin images from the 2019 ISIC Challenge dataset [9]. We designed and implemented the classifiers as CNNs using TensorFlow [18], and planned on reducing anticipated overfitting by expanding training images using an image augmentation technique.

2 Related Work

Over the years, new techniques that seek to improve the accuracy of skin cancer diagnosis and detection have emerged in the biomedical world. In 2012, Sheha et. al. [16] developed a system based on the use of two different multilayer perceptron (MLP) classifiers. The highest classification accuracy was achieved by using the traditional MLP classifier over an automatic one, with 92% and 76% classification accuracies, respectively [18].

Sharma and Srivastava [15] have proposed the use of Back-propagation neural networks (BNNs) and Auto-associative neural networks (AANNs) for accurate prevention and diagnosis of skin cancers. They developed a BNN-based classifier that achieved a 90.2% overall accuracy with three hidden layers of 40, 25 and 10 neurons in each hidden layer. In this case, the high number of neurons per hidden layer helped reduce the probability of overfitting. They also developed an AANN-based classifier that achieved overall accuracy of 81.5

It should be stated that the type of image you work with in some cases affects the performance of a certain diagnostic method [20]. In the past it has been observed that using dermoscopic image sets improves the accuracy with which CNNs classify malignant cases, while CNNs that were trained with close-up images were better at diagnosing benign cases [20]. Dermoscopic images are obtained through dermoscopy, an imaging technique which removes the reflection of the skin on the surface [8]. Thus, compared to close-up images, dermoscopic imaging provides an enhanced visualization of the skin [8].

3 Data Preparation

The 2019 ISIC Challenge dataset [10] contains 25,331 dermoscopic images for training with labels across 8 different categories. A part of the dataset is the HAM10000 dataset. Of these images, we use 80% as a training set and the rest as a validation set. All the images in the dataset come with different sizes, so we apply random cropping with a fixed resolution of 256x256 pixels. Random cropping allows our models to generalize better, for what we want them to learn is not always completely visible in the images or the same scale in our training data [17]. We also applied image augmentation technique to artificially expand the dataset and to increase the accuracy with fewer images. Specifically, we use some properties of the ImageDataGenerator class of Keras, such as rescale, rotation, shear range and zoom range for image augmentation. During the whole training, we had to deal with the fact that the initial dataset was heavily imbalanced, and the number of images in some classes was greater than in the rest, as reported in Table 1. For instance, for the Melanocytic Nevus class we have 12,875 images, almost half of the entire dataset, while the class Dermatofibroma consists of just 239 images. Precisely, four of these categories constitute 92% of the available data, while the rest make up only 8%. The ratio of malignant and benign images is about 1 to 2 in the training dataset. We organized the image data in subfolders according to the type of skin cancer or lesion to which they belong, so that the ImageDataGenerator class can make use of it.

Table 1: Number of Samples for Each Class in the Training Dataset

Diagnostic Class	Number of Images	Percentage
Melanoma - MEL	4522	17.9%
Melanocytic Nevus - NV	12875	50.8%
Basal Cell Carcinoma - BCC	3323	13.1%
Actinic Keratosis - AK	867	3.5%
Benign Keratosis - BKL	2624	10.4%
Dermatofibroma - DF	239	0.9%
Vascular Lesion - VASC	253	1%
Squamos Cell Carcinoma - SCC	628	2.4%
Total	25,331	/

4 Training and Testing

We built two models - a binary classifier for malignant and benign cases, and a multiclass classifier to classify the lesions and types of skin cancer reported in Table 1. Initially, we developed and trained our models on a local computer in Jupyter Notebook [11] environment, but the time to train the model by going through the whole training set just one time, also known as an epoch, was too long due to lack of computing power. Usually training a model takes well more than a hundred epochs, so training models on a local computer turned out to be an infeasible solution. Then, we updated our models and moved the development environment to Google Colaboratory [5] where users can develop and execute Python code in Jupyter Notebook environment with cloud computing free of charge. Still, the code in Google Colaboratory couldn' t run for more than about 30 minutes due to timeout.

As a consequence of receiving the Google Cloud Platform (GCP) Research Grant, we were able to set up virtual machines with multiple GPUs and train without time limit. This way, the time to run a single epoch went from about three and a half hours on a local machine to 4 minutes on a virtual machine with GPUs.

All training and tests ran using different configurations of GPUs to be able to decrease the time to run epochs as much as possible. For our binary classifier, we initially used a virtual machine with two NVIDIA Tesla T4 GPUs with 64GB of video memory, but in order to run more epochs in less time we took a step forward and upgraded to eight NVIDIA Tesla K80 GPUs with 128GB of video memory. For our multiclass classifier we used a virtual machine with two NVIDIA Tesla T4 GPUs and 64GB of video memory, without upgrading. In the multiclass classifier, the upgrade was unnecessary because we didn' t' need to train as many epochs since overfitting happened at an earlier epoch.

Convolutional Neural Networks (CNNs) emerged from the study of the brain' s visual cortex and are known to perform well in computer vision tasks such as image classification [6]. A CNN can be defined as a Deep Learning model which can analyze images, define weights and biases for different objects in those images, and be able to distinguish one from another. CNNs are made up of different layers, of which the convolution and pooling layers are very important. The convolution layers transform the input image so as to extract features from it. The pooling layers are used to reduce the size of the input image.

So as to achieve the best training model, we tried different configurations, changing the number of neurons per layer and the number of convolutional / pooling layers. The models that are presented are the ones that have obtained the best performances. Our CNN models are composed of a Sequential layer that includes four pairs of convolutional and pooling layers, a Flatten layer that

shapes 2d image data into single dimensional data, and a Dense layer which consists of densely connected layers. There is a Dropout layer, which applies the dropout regularization technique by randomly disabling neurons at certain percent during training to avoid overfitting in artificial neural networks, just before the dense layer. The dropout layer has a 50

Regarding the number of filters per convolution layer, we tried different combinations to see how the model would respond. There is no standardized way to choose that, so most of the time the combination depends on the data available and the results obtained. In the end, the number of filters in the four convolutional layers set to 32, 64, 64, and 128 obtained the best results. The configuration of our CNN is shown in Figure 1.

The Adam optimizer, which is known to converge to the global optimum quickly, is one of the most popular optimizers in deep learning [6], and was used for training our models. The last layer of the model is an output layer with the SoftMax activation function for our multi-class classifier, or with the Sigmoid activation function for our binary classifier. The design of the models is shown in Figure 1.

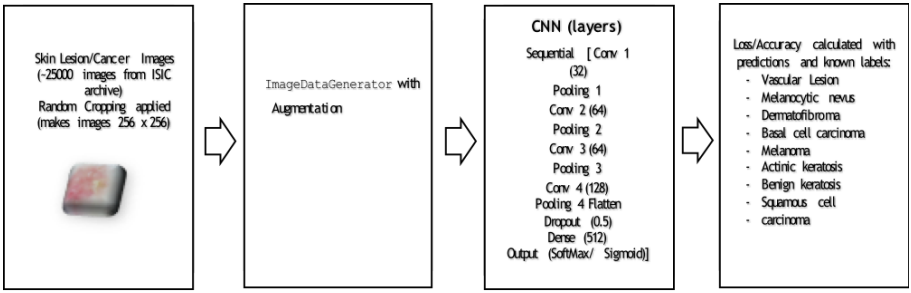

Figure 1: The Training Process of the Skin Cancer Classifier

With the training and test image data ready, we trained the models with the data, considering the loss and the accuracy, which are indicators of how the training process is going. We used a split ratio of 80:20 of training and validation data for both our binary classifier and multiclass classifier. After each epoch, the accuracy and loss of the models were logged and graphed. Accuracy is the portion of correct predictions, and loss is a measure of a model' s performance. We use the Cross-Entropy loss function for our models. More precisely, the Binary Cross-Entropy loss function for our binary classifier, and the Categorical Cross-Entropy one for out multiclass classifier. Both loss functions measure the performance of a model whose output is a probability value between 0 and 1 [13]. We have to use slightly different Cross-Entropy loss

functions because, based on the nature of our classifiers, we may have more than two classes. We run batches of 25 epochs to account for overfitting and we save it every 50 epochs of training to prevent data loss due to a connection timeout. After each epoch, training accuracy and validation accuracy are recorded. Training accuracy is the percent of correct predictions by the model being trained on the training data, and validation accuracy is calculated from the validation data. We use the validation set to compare the two obtained accuracies in order to correctly evaluate the results and observe the performance of the classifier; Figures 2 and 3 show training and validation accuracy curves of the models as training takes place.

Training has to stop when overfitting happens. Overfitting occurs when the model performs well on training data, but poorly on data that were never used during the training. We consider the occurrence of overfitting as the gap between training accuracy and validation accuracy gets bigger. In Figure 2, approximately at 100 epochs, overfitting is observed with the training accuracy at about 72%. In Figure 3, overfitting is not observed, even though the accuracy is plateaued at 79% at around 200 epochs.

5 Result

Our binary classifier' s accuracy approaches 79% with 220 epochs of training, and our multiclass classifier' s accuracy approaches 72% with 100 epochs, as described in Table 2. Evaluating the accuracy curves for each model, we observed that the multiclass classifier began to overfit around at 100 epochs, while the binary classifier performed well without overfitting until it plateaued, as shown in Figures 2 and 3.

Table 2: Classifier Accuracy

Classifier Type	Validation Accuracy
Binary (Malignant or not)	79%
Multiclass (Types of skin cancer/lesion)	72%

Figures 4 and 5 show the confusion matrix, and present the prevalence of misclassifications on each label for the classifiers. Basically, the bright blocks in the diagonal show high accuracy, and MEL, BCC, BKL and NV show high accuracy. The number of images per diagnostic category in Table 1 helps us understand why some of these present more misclassifications than others. As mentioned above, the initial dataset was heavily imbalanced, therefore affecting those diagnostic categories with fewer images. As a matter of fact, NV,

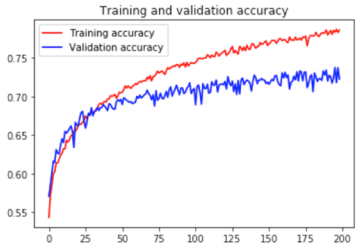

Figure 2: Accuracy Curve of the Skin Cancer Multiclass Classifier.

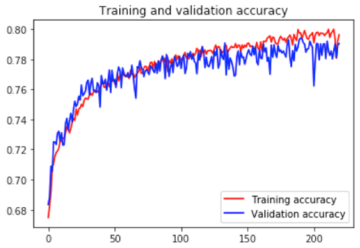

Figure 3: Accuracy Curve of the Skin Cancer Binary Classifier.

BCC, MEL, and BLK were likely to be classified correctly than others in the multiclass classifier due to the greater number of images in the training set.

Observing Figure 4 and 5, we notice from the confusion matrices that the number of true positives is quite low compared to the true negatives. This is another consequence of the heavily imbalanced training set and validation set. The ratio of images representing malignant skin cancers/lesions to benign ones is 8,473 to 17,058, roughly 1 to 2.

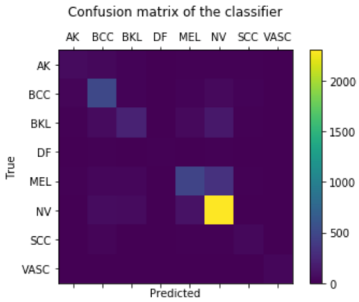

Figure 4: Confusion Matrix of the Multiclass Classifier

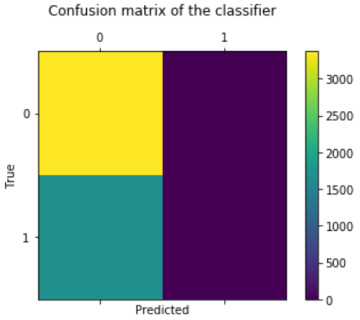

Figure 5: Confusion Matrix of the Binary Classifier

6 Conclusion

In this paper, we have presented a deep neural network using CNNs that we designed and developed to detect skin cancer and lesions. The results are promising with an accuracy of 72% (multiclass classifier) and 78% (binary classifier). It also confirmed that the accuracy of the model is heavily influenced by the size and ratio of the classes in the training set.

Future studies exploring new learning models, such as transfer learning that trains an existing well-performing model rather than from the scratch, may be needed to improve accuracy. A focus on improving the pre-processing, thus making it more effective so as to handle the imbalanced data during the skin lesion/cancer analysis, is another area of future research.

We believe research like ours plays an important role in the early detection and treatment of skin cancers thanks to its simplicity and accuracy. Furthermore, this kind of research can aid in developing other tools, such as apps for physicians and patients to use, further decreasing the time required to arrive at a correct diagnosis and, subsequently, expediting the treatment of what can be a debilitating disease.

References

[1] S. A. AlShehri and S. Khatun. Uwb imaging for breastr cancer detection using neural network. *Progress In Electromagnetics Research C*, 7:79–93, 2009. doi:10.2528/PIERC09031202, Last accessed on 2020-04-11.

[2] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B. Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019. doi:10.1007/978-3-030-32245-8_11, Last accessed on 2020-05-24.

[3] Sam Blanchard. Cancer is 'overtaking heart disease as wealthy countries' biggest killer'., 2019. "https://www.dailymail.co.uk/health/article-7421917/Cancer-overtaking-heart-disease-wealthy-countries-biggest-killer.html", Last accessed on 2020-02-10.

[4] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, Wei Xie, Gail L. Rosen, Benjamin J. Lengerich, Johnny Israeli, Jack Lanchantin, Stephen

Woloszynek, Anne E. Carpenter, Avanti Shrikumar, inbo Xu, Evan M. Cofer, Christopher A. Lavender, Srinivas C. Turaga, Amr M. Alexandari, Zhiyong Lu, David J. Harris, Dave DeCaprio, Yanjun Qi, Anshul Kundaje, Yifan Peng, Laura K. Wiley, Marwin H. S. Segler, Simina M. Boca, S. Joshua Swamidass, Austin Huang, Anthony Gitter, and Casey S. Greene. Opportunities and obstacles for deep learning in biology and medicine. *The Royal Society*, 2018. doi:10.1098/rsif.2017.0387, Last accessed on 2020-03-19.

[5] Google. Google colaboratory. "https://colab.research.google.com/notebooks/intro.ipynb", Last accessed on 2020-07-09.

[6] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. OReilly, 2019. Last accessed on 2020-06-16.

[7] Xiao Han. Mrbased synthetic ct generation using a deep convolutional neural network method. *Medical Physics Journal*, 2017. doi:10.1002/mp.12155, Last accessed on 2020-02-20.

[8] Zalaudek I., Argenziano G., Di Stefani A., Ferrara G., Marghoob A.A., Hofmann-Wellenhof R., Soyer H.P., Braun R., and Kerl H. Dermoscopy in general dermatology. *Karger*, 212:7–18, 2006. doi:10.1159/000089015, Last accessed on 2020-06-19.

[9] ISIC. Isic 2019. "https://challenge2019.isic-archive.com", Last accessed on 2020-04-15.

[10] ISIC. Training data. "https://challenge2019.isic-archive.com/data.html", Last accessed on 2020-02-10.

[11] Project Jupyter. Project jupyter. "https://jupyter.org/about", Last accessed on 2020-07-09.

[12] June-Goo Lee, anghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and NamkugS Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 2017. doi:10.3348/kjr.2017.18.4.570, Last accessed on 2020-03-12.

[13] Nielsen and Michael A. Improving the way neural networks learn. "https://neuralnetworksanddeeplearning.com/chap3.html", Last accessed on 2020-05-18.

[14] American Academy of Dermatology. Skin cancer. "https://www.aad.org/media/stats/conditions/skin-cancer", Last accessed on 2020-02-10.

[15] Deepti Sharma and Swati Srivastava. Automatically detection of skin cancer by classification of neural network. *International Journal of Engineering and Technical Research*, 4:15–18, 2016. issn:2321-0869, Last accessed on 2020-05-15.

[16] Mariam A. Sheha, Mai S.Mabrouk, and Amr Sharawy. Automatic detection of melanoma skin cancer using texture analysis. *International Journal of Computer Applications*, 42:22–26, 2012. doi:10.5120/5817-8129, Last accessed on 2020-05-15.

[17] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *Journal of LaTeXClass Files*, 14:1–11, 2015. "https://arxiv.org/pdf/1811.09030v1.pdf", Last accessed on 2020-06-19.

[18] TensorFlow. Tensorflow. "https://www.tensorflow.org/", Last accessed on 2020-07-08.

[19] T. T. P. Thanh, Caleb Vununu, Sukhrob Atoev, Suk-Hwan Lee, , and Ki-Ryong Kwon. Leukemia blood cell image classification using convolutional neural network. *Internation Journal of Computer Theory and Engineering*, 10:54–58, 2018. doi:10.7763/IJCTE.2018.V10.1198, Last accessed on 2020-06-13.

[20] Philipp Tschandl, Cliff Rosendahl, and Bengu Nisa Akay. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatology*, 155:58–65, 2019. doi:10.1001/jamadermatol.2018.4378, Last accessed on 2020-04-12.

[21] Richard C. Wender. Barriers to effective skin cancer detection. *American Cancer Society Journals*, 1995. "https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.1002/1097-0142%2819950115%2975%3A2%2B%3C691%3A%3AAID-CNCR2820751412%3E3.0.CO%3B2-G", Last accessed on 2020-03-12.