Skin Lesion Classification with CNN and Physical Layers

Michelle Li
Department of Biomedical Engineering
Duke University
Durham, NC 27708
mjl103@duke.edu

Charlotte Roh
Department of Electrical & Computer Engineering
Duke University
Durham, NC 27708
hr67@duke.edu

Yuncheng Duan
Department of Biology
Duke University
Durham, NC 27708
yd90@duke.edu

Roger Chang
Department of Biomedical Engineering
Duke University
Durham, NC 27708
yc438@duke.edu

Abstract

Skin cancer is the most prevalent type of cancer, and can be deadly if left untreated. In recent years, deep learning approaches have been implemented to automate skin cancer detection and classification using dermoscopic images, however, the process still requires human intervention to capture accurate and high quality dermoscopic images. This study aims to automate the physical components of the dermoscope to see if optimizing for these components will increase model performance results. After optimizing for the color channels, illumination patterns, and lens blur with a convolutional neural network (CNN), the results suggest training the lens blur with the CNN improves classification performance, from 0.70 AUC and 70% accuracy to 0.73 AUC and 74% accuracy. This study can be extended to a larger aim of creating inexpensive, portable and fully automated imaging systems to detect skin cancer.

1 Introduction

Skin cancer is outlooked to affect more and more people every year. Melanoma, specifically, is responsible for 75% of skin cancer deaths, despite being the least common skin cancer [1]. In the status quo, diagnosis is done manually by healthcare workers. However, as with other cancers, early and accurate detection—potentially aided by data science—can make treatment more effective.

Existing artificial intelligence approaches have not adequately considered this clinical frame of reference. Dermatologists could enhance their diagnostic accuracy if detection algorithms take into account "contextual" images within the same patient to determine which images represent melanoma cancer. If successful, classifiers would be more accurate and could better support dermatological clinic work.

2 Related Work

Melanoma identification and diagnosis through machine learning is a problem that has been tackled before. The context of many published works pertaining to melanoma diagnosis highlights the pertinence of these models of healthcare, alluding to future mobile applications that will help detect and diagnose skin cancer. Many of the first published models for melanoma classification focused on

segmentation masks, clustering, and supervised learning [2]. The commonality between these models is the importance of border, diameter, and asymmetry for feature extraction [2, 3]. The complexity of the models used in these papers is directly proportional to the scale of the classification task. When only performing binary classification on melanomas versus all other types of skin lesions, Visual Geometry Group 16 (VGG16) or AlexNet are commonly used as the base architecture for the model. However, with larger classification tasks that branch out to classify subsets of benign, malignant, and non-neoplastic lesions, more complex networks like DenseNet or Google's Inception v3 CNN are used for their expressive power [4].

In more recent works, the performance of these models is compared directly to the diagnoses made by dermatologists when looking at the images with the raw eye [4]. Models have been shown to have higher accuracy while dermatologists' accuracy is reduced by false positive rates [4]. While these models generally have high accuracy and low loss, the most accurate diagnosis for skin cancer is by performing a biopsy. Many of the models discussed and cited here were built to illuminate the possibility of cancer as a push for early detection.

Inspiration for the physical layer used in this experiment came from a variety of experiments, especially those simulating a phase-coded aperture physical layer before their model [5]. These experiments show that allowing a model to determine the most optimal state for the feature being studied (blur in this specific case) will generally result in better performance. The physical layers were automated and the methods are described below.

3 Methods

3.1 Dataset

The dataset used for this project was extracted from the International Skin Imaging Collaboration Challenge (ISIC 2020), which contains dermoscopic images from a variety of hospitals including Hospital Clínic de Barcelona, Medical University of Vienna, etc. [6]. The dataset contains 33,126 dermoscopic images of skin lesions and their respective ground truth labels of being benign or malignant, from over 2,000 patients [6]. The dataset also includes metadata entries of patient ID, sex, age, and general anatomic site. 300 benign and 300 malignant images, for a total of 600 images, were chosen for this study. 70% of the data was used for training, 15% for validating, and 15% for testing. Due to the large image files and constraints in Google CoLab, only a small subset of the total ISIC 2020 images were used, and the images were resized to 256x256 for inputs into our models. To make up for the small training data and attempt to reduce overfitting, data augmentation techniques were initially implemented on the training data, such as rotation, as well as horizontal and vertical flips, but that did not improve performance results for our baseline model. Thus, training data remained not augmented for the rest of the models.

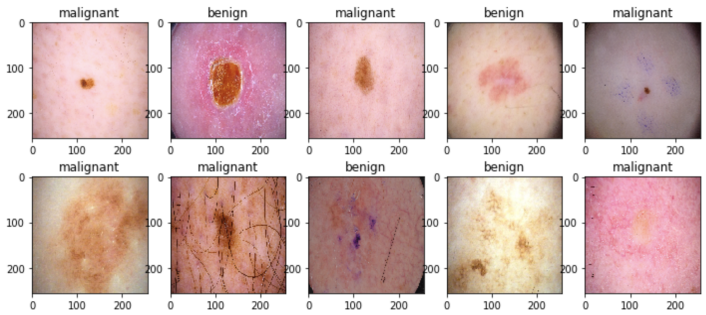


Figure 1: Skin lesion dermoscopic images their classification labels

Figure 1 shows some examples of the dermoscopic images from the ISIC 2020 dataset with their respective classification labels. The images will be used as inputs in our models and the labels will be used as ground truths for classification.

3.2 Exploratory Data Analysis

An exploratory data analysis has been conducted to truncate unnecessary features and to get a general overview of the dataset.

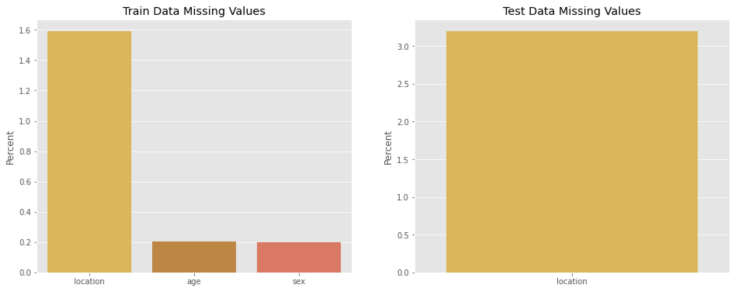


Figure 2: Features that have missing values in train and test dataset

In order to process the data, first empty values in both train and test datasets had to be checked as shown in Figure 2, because they will throw the model off. There were different approaches to different data types: for a numerical column, the NaN values were filled with the mean of the column, and for categorical data, the column was filled in with another value called 'Missing'. However, for columns that contain boolean values, the column is filled with whatever value is the majority in the column. That way, all NaN values were gone from the dataset.

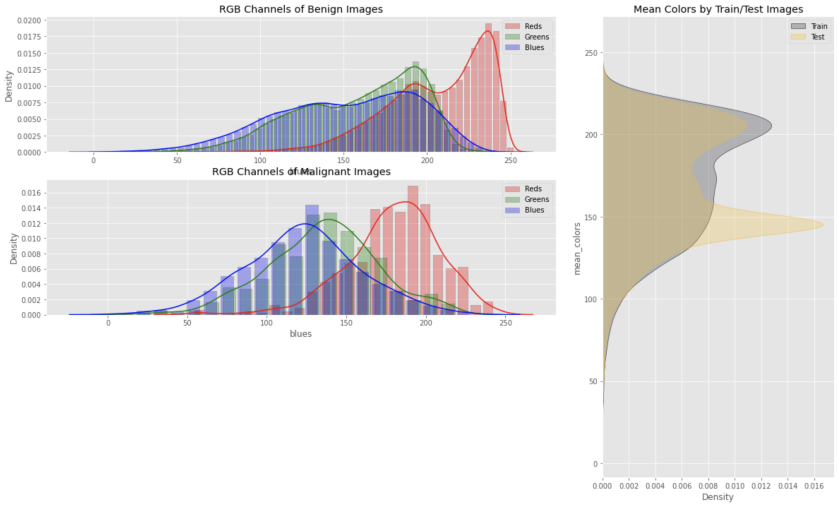


Figure 3: RGB color channel distributions in both training and test images

Figure 3 shows the color channel distributions in both training and testing images. The distribution discrepancies for the images suggest that this color feature may be a prominent feature in classifying for different classes, thus eventually a physical layer as described in Section 3.3 has been designed based on this feature.

In training models, some algorithms demands the absence of collinearity, as if two certain features have a high rate of collinearity, it would not yield a significant prediction rate. Therefore, eliminating highly correlated features is essential to avoid potential pitfalls. A correlation analysis was conducted heatmap visualization illustrated in Figure 4, which highlights pairs of features with high correlation coefficient. Table 2 outlines the correlation plots for each feature. A high correlation is observed between 'age', 'image_size', 'width', and 'height'. Thus, it was a natural conclusion that all of those columns were dropped.

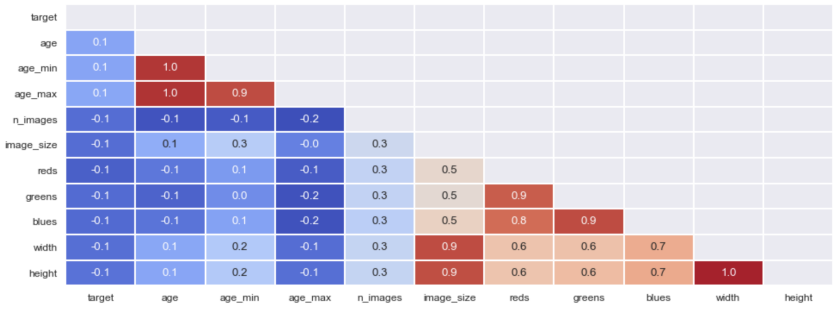


Figure 4: Correlation plot between feature pairs

3.3 Training

The baseline model is a convolutional neural network, and the architecture is shown in Figure 5. The CNN architecture contains 4 convolutional layers, 2 dense layers, and 2 max pooling layers.

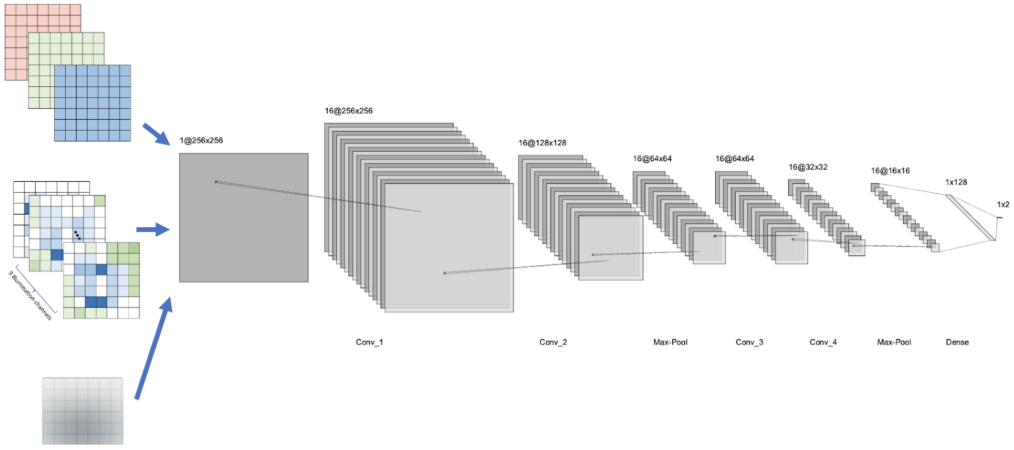


Figure 5: CNN Model Architecture

To simulate the physical components of our imaging system, physical layers were added into the CNN and trained with the models. The physical layers include RGB color channels, illumination pattern, and lens blur. For RGB channels, each image was split into red, green and blue color channels. Three trainable parameters were assigned to the RGB layers as the weights of the three colors. 1-1 convolution was used to add the trainable weights into the model. For illumination patterns, nine illumination phase angles were included into the analysis (a combination of -10, 0, 10 from x and y axes). The illumination patterns depended on the optical thickness of each part in the image. First the color images were converted into the gray scale images, and then light waves from different angles were applied to the images by simulation. Nine illumination channels per image were obtained after this procedure. Similar to the analysis on the RGB channels, nine trainable weights were assigned to the illumination channels and those weights were included into the CNN model for training. As for lens blur, Gaussian kernels with five different kernel sizes, namely 10, 25, 50, 100, and 200, were applied on each image. Images with the five Gaussian kernels were sent as the input to train the models respectively. No trainable parameters were added to the kernels in this step.

The model performances were evaluated using accuracy, precision, recall, AUC, and F1 score.

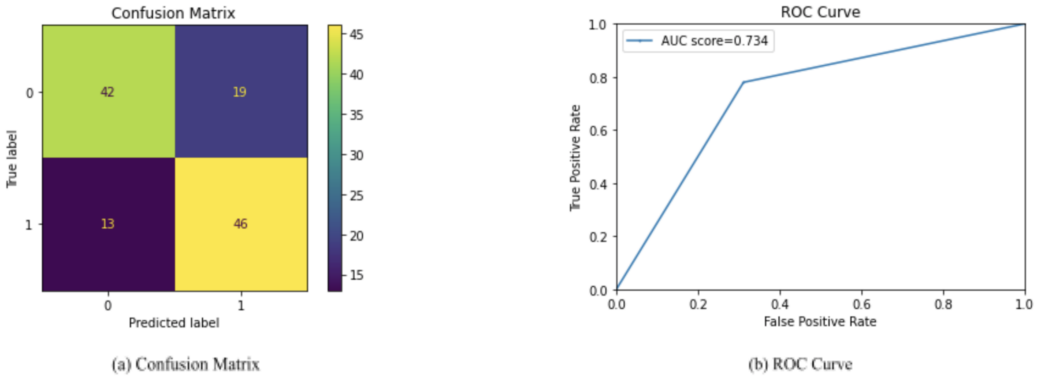
4 Results

Table 1 shows the performance results of the four different models evaluated using the test set. The accuracy of all the models fall in the range of near 65% to 75%. Since accuracy is not the best metric to determine performance, confusion matrices were created for each model to see the result of the classification, as well as to evaluate each model based on precision, recall, F1-score, and AUC score.

Table 1: Quantitative Performance Results for Different Models

	Test Accuracy (%)	Precision	Recall	F1-score	AUC
CNN	70	0.67	0.80	0.73	0.70
RGB	64	0.63	0.66	0.64	0.63
Illumination	68	0.65	0.73	0.69	0.65
Gaussian	74	0.69	0.88	0.78	0.73

For the RGB model, the optimized weights were 0.04, -0.03, and -0.10, respectively. The blue channel was weighted more than the rest of the color channels.



(a) Confusion Matrix (b) ROC Curve

Figure 6: ROC curve and the confusion matrix of the model with the Gaussian kernel layer

The confusion matrix and the ROC curve of our best model is shown in Figure 6, which is the Gaussian lens blur model with a kernel size of 10. Figure 6 (a) shows that the number of false-positive samples is high, resulting in a low precision score of 0.69. As seen in the ROC curve in Figure 6 (b), the AUC value is 0.73, which is not perfect, but still suggests that the model can classify skin lesions effectively to a certain degree.

5 Discussion

Based on the results, the precision scores for each model is very similar, suggesting that the ability for each model to identify the real positive samples (the malignant skin lesions) among all predicted positive samples is quite similar. These results imply that our classifiers are not great enough to learn features that differentiate benign skin lesions from malignant skin lesions. However, the simple CNN model and the model with the Gaussian kernel model have the highest recall and F1-scores, representing that they have better abilities to predict the positive samples among all real positive samples. The blur model may have performed the best because adding blur can help highlight important features in the images and blur out unimportant artifacts, such ruler markings, discoloration, and hair. Only 1 Given more time, trainable weight parameters would be added to optimize for the best blur.

Limitations to this study include the limited memory space in Google Colab, which is why a larger dataset to train the models was not used. Since the baseline CNN performance was mediocre, a more complicated CNN architecture, ResNet50, was initially implemented, but performed worse than the

simple CNN model due to overfit from many parameters but little training data. While max pooling and dropout layers were implemented, all the models overfit the data during training.

For future improvements, the overfit problem will be tackled by trying better data augmentation techniques or training with more images. In addition, more state-of-the-art classification models can be used, such as FixRes from Facebook's AI team or EfficientNet from Google's AI team, to classify the lesions. Since certain illumination phase angles and Gaussian blur kernels were manually chosen, the optimal phase patterns and lens blur for classification may not be found yet. Thus, adding more parameters to the physical layers, or even training different types of physical layers together in one model, are possible ways to improve classification performance.

Other future directions include testing how image segmentation performance tasks are affected by the physical components of the imaging system, as well as training the models on other types of biomedical imaging data such as cell images or CT scans to see how robust the models are. Since metadata information was also provided, integrating both metadata and images into the same model can result in improved malignant skin lesion prediction performance.

6 Conclusion

Overall, it can be concluded that the lens blur model was the only model that outperformed the baseline CNN on all evaluation metrics. The illumination model was the second best performing model, followed by the RGB model. This study suggests that the current imaging system can be improved to create more efficient, portable and inexpensive biomedical imaging devices.

References

[1] Pushkar Aggarwal, Peter Knabel, Alan B. Fleischer, "United States burden of melanoma and non-melanoma skin cancer from 1990 to 2019," Journal of the American Academy of Dermatology, Volume 85, Issue 2, 388-395, 0190-9622 (2021).

[2] Yuexiang Li and Linlin Shen. Skin Lesion Analysis Towards Melanoma Detection Using Deep Learning Network, 2017; arXiv:1703.00577

[3] Aminur Rab Ratul, M. Hamed Mozaffari, Won-Sook Lee, and Enea Parimbelli. Skin Lesions Classification Using Deep Learning Based on Dilated Convolution, 2019.

[4] Amirreza Rezvantalab and Habib Safigholi and Somayeh Karimijeshni. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms, 2018; arXiv:1810.10348.

[5] Shay Elmalem, Raja Giryes, and Emanuel Marom. Learned phase coded aperture for the benefit of depth of field extension, 2018.

[6] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., et al. "A patient-centric dataset of images and metadata for identifying melanomas using clinical context". Sci Data 8, 34 (2021).

[7] Wei Li, Alex Noel Joseph Raj, Tardi Tjahjadi, Zhemin Zhuang, "Digital hair removal by deep learning for skin lesion segmentation," Pattern Recognition, Volume 117, 107994, ISSN 0031-3203 (2021).

[8] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).