A Comparison of the Efficacy of Image Classification via Convolutional Neural Networks and Logistic Regression

University of North Carolina at Chapel Hill
Mattheus Bezerra, Eric Ashton, Blaine Hill, Samuel Friedman
COMP562 Spring 2021
Professor Jorge Silva

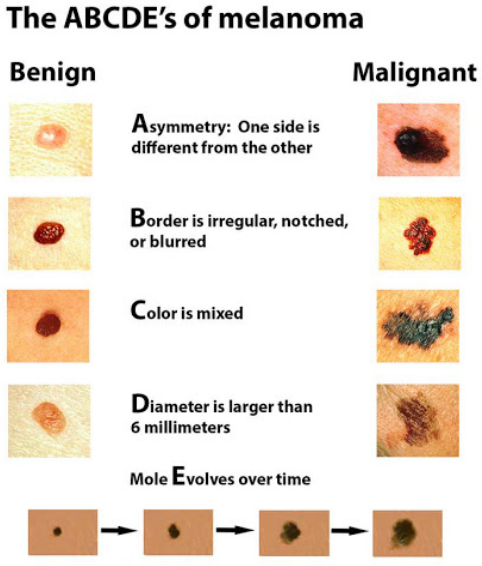
Problem Description

Melanoma is a highly aggressive cancer that develops in melanocytes, the cells that produce melanin. It can also form in the eyes and, rarely, inside openings of the body such as the nose or throat. It can be detected via irregular lesion shape, size, changes over time and color. Melanoma can be harmless or lead to cancer[1]. It is much less common than some other types of skin cancers but is more dangerous due to its potential to spread to other parts of the body if not caught and treated early[2]. Being able to identify early-onset melanoma is crucial to dermatologists and their patients and can save thousands of lives.

Today, there is not a universally acclaimed model in practice for identifying melanoma. Therefore, Kaggle held a competition on their site requiring participants to use a given dataset to identify the difference between malignant and benign skin lesions[3]. With machine learning tools and algorithms like convolutional neural networks, dermatologists are greatly assisted in the process of diagnosing the disease early. Doctors can then treat the malignant melanoma if it is correctly identified early. This project utilizes the same Kaggle dataset to train and test the melanoma image data using a convolutional neural network (CNN).

Hypotheses and Research Questions

The purpose of both models is to classify skin lesions as benign or cancerous. An accurate model to classify lesions could augment current, mathematically-based predictive methods and improve diagnosis or research efforts. This model has the potential to identify tumors early and decrease preventable deaths due to late diagnoses.



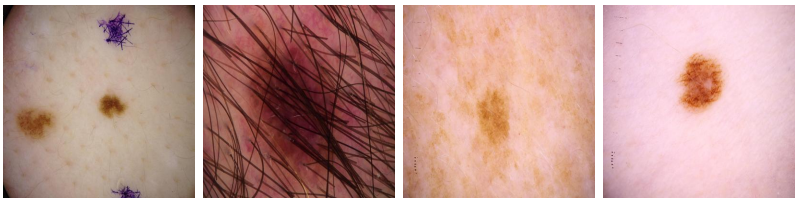
Summary of the Harvard Health Publishing's symptoms of Melanoma[4] to the left

While gathering data and considering potential models, the following questions were explored:

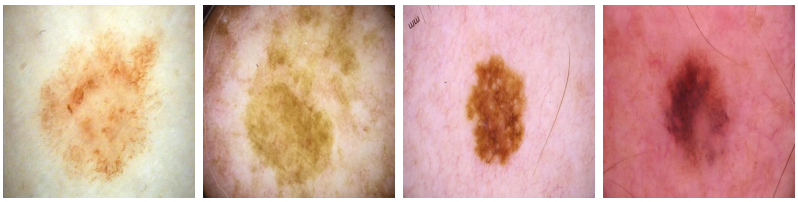
1. Is the severity of a skin lesion predictable by machine learning and computer vision models?
2. What predictive model is most effective in classifying images of malignant and benign lesions?

Data and Data Sources

The dataset contains 33,126 dermoscopic training images of unique benign and malignant skin lesions from over 2,000 patients. Each image is associated with one of these individuals using a unique patient identifier. All malignant diagnoses have been confirmed via histopathology, and benign diagnoses have been confirmed using either expert agreement, longitudinal follow-up, or histopathology. A thorough publication describing all features of this dataset is available in the form of a pre-print that has not yet undergone peer review[5]. The public training dataset contains a split of 32,542 benign skin lesions and 584 malignant skin lesions.



Benign - these represent melanoma that is most likely not harmful



Malignant - these would be classified as high-priority for treatment

The testing set for this dataset is available on Kaggle without any labels. Due to the size of the dataset and lack of computational resources, 2526 images were used for training and 510 for testing. All malignant class images were used. The images were resized to be 100x100 using MATLAB. However, all data was transformed via flipping both horizontally and vertically to provide enough images for the models to train later. There were a total of 7578 images for training and 1530 images for testing after these transformations. In the case of the CNN, a separate validation set was created during training using scikitlearn's train-test-split function. The validation split comprised 15% of the sampled training data. Both models were run via Google Colab's GPU.

The dataset was generated by the International Skin Imaging Collaboration (ISIC), and images are from the following sources: Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, The University of Queensland, and the University of Athens Medical School. The dataset was curated for the **SIIM-ISIC Melanoma Classification Challenge** hosted on Kaggle during the Summer of 2020.

Model Description

Two machine learning models were utilized in an effort to classify benign and malignant skin lesions. The first model was a simple logistic regression used as a control to compare its results to the second model: a CNN.

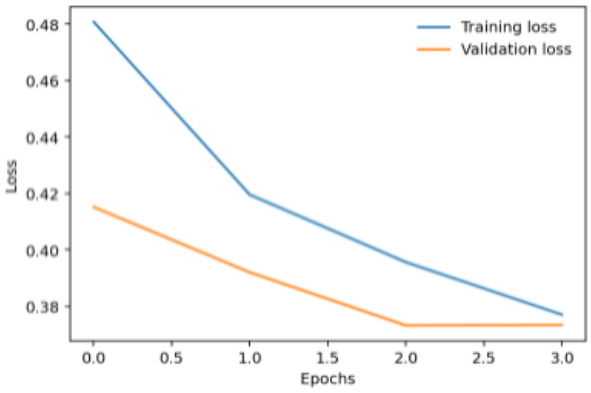
The CNN was constructed using PyTorch and consists of two convolutional layers, followed by 4 dense layers with one layer of dropout. Each dense (Linear) layer contains a Leaky ReLU activation function. The full architecture of the CNN is provided below.

```
(conv1): Conv2d(3, 5, kernel_size=(5, 5), stride=(1, 1))
(conv2): Conv2d(5, 5, kernel_size=(5, 5), stride=(1, 1))
(conv2_drop): Dropout2d(p=0.15, inplace=False)
(fc1): Linear(in_features=2420, out_features=1024, bias=True)
(fc2): Linear(in_features=1024, out_features=512, bias=True)
(fc3): Linear(in_features=512, out_features=150, bias=True)
(fc4): Linear(in_features=150, out_features=2, bias=True)
```

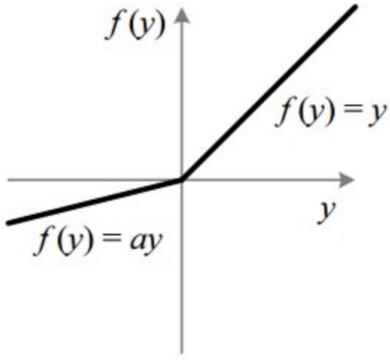
Results and Discussion

The logistic regression model was designed using SciKit Learn. The model performed with 83.7% accuracy on the training data with a testing accuracy of 81%. Its final positive predictive value (PPV) 22.6% with a recall of 15.6%.

The CNN performed with a training accuracy of 83.1% testing accuracy of 87.4%. Loss consistently decreased for training. Its final PPV was 46.6% with a recall of 50%. The following graphs, respectively from left to right, depict the model loss through training. The model's training included preset seeds to produce consistently deterministic results. This allowed us to graph loss to identify a divergence between training and validation loss which would indicate overfitting.

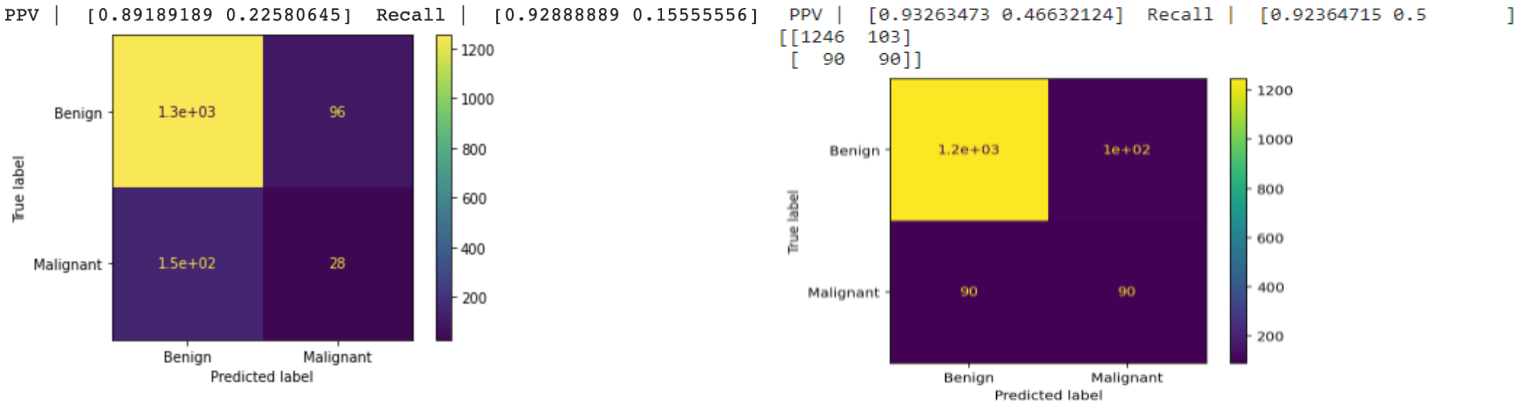


Model Loss



Leaky ReLU activation function

The SciKit Learn logistic regression confusion matrix is depicted on the left, and the CNN confusion matrix is shown on the right. These images also display the PPV, also commonly known as precision, and recall metrics. Note that these metrics are calculated on both the benign and malignant classes. In a clinical setting, the repercussions of a false negative diagnosis can be fatal. Thus, we focus on the metrics using the true positives. Recall is the rate of true positives TP / (FN + TP). PPV is the rate of TP / (TP + FP).



The accuracy could be improved in the future by using more training data. The minority (malignant) class composed <2% of the entire data set, implying that further positive samples may improve predictive metrics. Several finalists from the Kaggle competition used advanced networks such as AlexNet, ResNet, and DenseNet. Perhaps a more sophisticated architecture or gridsearch to optimize the hyperparameters of this model could improve performance. For example one may adjust the training epochs, batch sizes, number of layers, layer sizes, dropout rate, learning rate, and so on. The data resizing may have also contributed to loss of information within the data. Using images at a larger or full size may allow for more complex information to be gathered and learned in another setting but were not feasible due to hardware limitations in Google Colab. The normalization of the images may also be adjusted to be better suited for highlighting skin lesions.

Although the CNN did not perform at a clinically viable level, it did expectedly outperform the logistic regression model by far. Firstly, the images are not uniformly centered around the skin lesions which may disrupt pixels being used as inputs versus convolutional layers being able to learn features. The images themselves are normalized, but they are highly variable in size, shape, and color. The problem may be too complex for a simple classification model. Alternatively, the CNN may have performed better due to its nonlinear layers, image convolution, and use of dropout to prevent overfitting. CNNs are traditionally used in image classification, so these results were predictable.

Ultimately, an improved version of our CNN model could be used to automate and assist the diagnosis of malignant melanoma. In conjunction with current, non-automated efforts, this could be an invaluable tool in catching as many cases as early as possible. Based on the final results, a signal does exist for classification of skin lesions but must be refined for use in a medical environment.

References

1) ***Melanoma - Symptoms and causes*. (2020, March 10). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884**

2) ***What Is Melanoma Skin Cancer? | What Is Melanoma?* (2021). Cancer.Org. https://www.cancer.org/cancer/melanoma-skin-cancer/about/what-is-melanoma.html**

3) ***SIIM-ISIC Melanoma Classification | Kaggle*. (2021). Kaggle.Com. https://www.kaggle.com/c/siim-isic-melanoma-classification/data**

4) **Harvard Health. (2020, June 17). *Melanoma*. https://www.health.harvard.edu/cancer/melanoma-overview**

5) **Rotemberg, V. (2021, March 5). *A patient-centric dataset of images and metadata for identifying melanomas using clinical context*. Scientific Data. https://www.nature.com/articles/s41597-021-00815-z**

Github: https://github.com/samuelfr98/Melanoma-Convolutional-Neural-Network