Melanoma Image Classification Using FastAI with the ResNet34 CNN

Caitlin Ortega Ruble

October 2022

Capstone Project for Springboard Data Science Career Track

Abstract

Melanoma is a deadly form of skin cancer that is highly curable when caught early. A deep-learning computer vision tool was developed to identify melanoma from images of skin lesions. This tool was trained on the SIIM-ISIC Melanoma Challenge dataset available on Kaggle using the FastAI library to interface with the ResNet34 pre-trained Convolutional Neural Network (CNN). The final model selected uses the average probability across 5 cross-validation folds to classify whether an image shows a malignant melanoma, or a benign skin lesion. In private validation testing, the ensembled model had a 90% chance of distinguishing between the positive and negative classes (AUROC = 0.9) and correctly classified 83% of the melanoma images (recall = 0.83). In further validation testing through submission to Kaggle, the model had an 85% chance of distinguishing between the malignant and benign classes (AUROC = 0.8531). A proposed use case for this model is as a screening tool for home use; early, easy, and skillful screening with artificial intelligence tools can effectively get more patients in the door for early treatment of melanoma. Getting more patients in the door helps melanoma patients by increasing their survival rates and decreasing the cost and intensity of treatment, hospital systems by reducing the number of patients in need of in-hospital surgery and chemotherapy and thereby reducing system strain, and health insurance companies by significantly reducing the cost of treating the same condition when caught early vs. late.

Introduction

Melanoma is a malignant type of skin cancer that is projected by the American Cancer Society to affect 100,000 new patients in the United States this year. Of those 100,000 diagnoses, 7,600 patients are expected to die from this virulent cancer. The good news is that melanoma is highly treatable when diagnosed early. In fact, the 5-year survival rate for patients diagnosed with melanoma while it is localized to the original region on the skin is 99%. These melanomas can usually be treated with simple in-office biopsy and excision procedures using local anesthesia. However, when the melanoma is not detected until it has spread through the region around the original source, the 5-year survival rate drops to 68%. The 5 year survival rate takes another steep hit if the melanoma is not diagnosed until it has spread throughout the body, dropping to just 30%. Applying skillful machine learning methods for early melanoma detection and diagnosis has the potential to positively impact patients, hospitals/doctors, and the insurance companies who foot the bill.

Approach

This computer vision problem called for a deep learning approach, and the use of a pretrained convolutional neural network (CNN) to be fine-tuned on the skin lesion data set was employed. The FastAI library coupled with the ResNet34 pretrained CNN was the selected approach. FastAI is a high-level framework built on top of PyTorch, designed to quickly create state-of-the-art results in neural network approaches. It allows for customization of how data is handled and model is trained, while utilizing pretrained models as a base architecture. The classes in the training images were balanced by down-sampling the benign images, and the resulting training subset was further split into a training and internal-validation set. The training data was manually split into 5 cross-validation folds, and the ResNet34 learner was trained separately 5 times over this cross-validation split. In each training, the ResNet34 model was fine-tuned over 15 epochs, optimizing for the binary RocAucScore metric. The average prediction value from all 5 CNN learners was used to return final prediction values for the internal validation set and the Kaggle-provided test set.

Data Set

The "SIIM-ISIC Melanoma Classification" dataset on Kaggle contains 33,126 labeled images and metadata for the 2,056 patients the images are taken from as training data. It contains an additional 10,982 unlabeled images with associated metadata as a test set. The images are available as DICOM files, JPG files, and TFRecord files. The metadata is available within the DICOM files and within the .csv files "train.csv" and "test.csv". The columns included in these .csv files were:

- image_name: unique identifier which points to the filename of the related image
- patient_id: unique patient identifier
- sex : the binary sex of each patient (male or female)
- age_approx: approximate age at the time of imaging, broken down into 5 year chunks
- anatom_site_general_challenge: location of imaged site
- diagnosis: detailed diagnosis information (train only)
- benign_malignant: classification label of each image, benign or malignant (train only)
- target: binarized version of the benign_malignant column (train only)

In addition to this dataset, the SIIM ISIC - 224x224 image set uploaded to Kaggle by user Arnaud Roussel were leveraged for model training. This dataset contained all the same imaged as the primary competition dataset, but had been resized to a standard 224x224 image size and saved as .png files. Using these resized images allowed for efficient model training relative to using FastAI to manually resize each image before being loaded for model training.

Exploratory Data Analysis

The training image dataset comprised of 33,126 unique images from 2,056 unique patients. The image dataset was very imbalanced, with only 1.8% of those images labeled as positive for the target class (Figure 1). Of the 2,056 patients included, 52% were male and 48% were female, showing an approximately equal split (Figure 2). The approximate age of the patients was normally distributed between 10 years of age and 90 years of age, with a mean age of 50 years and standard deviation of 15 years (Figure 3). The number of images included from each patient was highly variable, showing a strongly right-tailed distribution (Figure 4), however the number of images per patient was not found to be related to the number of malignant images per patient (Figure 5).

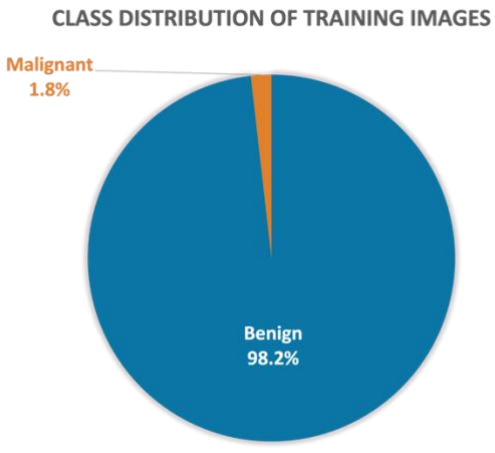

Figure 1: Strong class imbalance in the training set, with just 1.8% of training images positive for the target class.

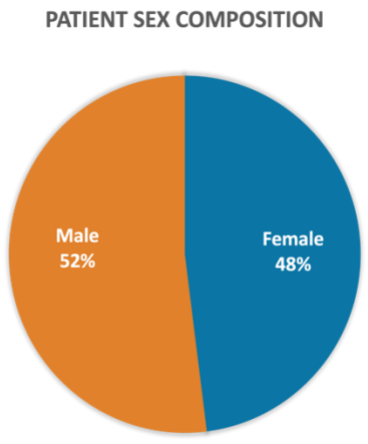

Figure 2: Male and female patients were included approximately equally.

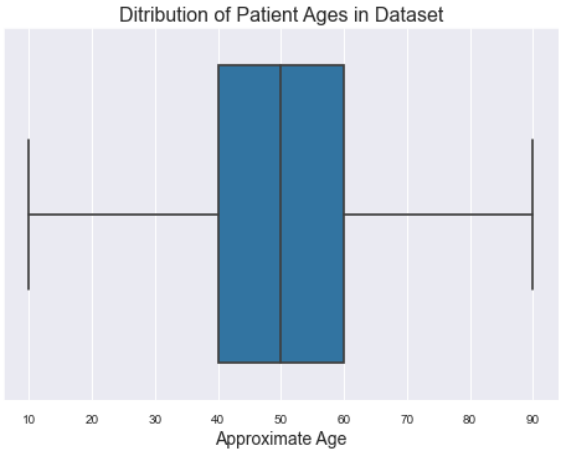

Figure 3: The ages of patients included in the study ranged from 10-90, with mean age 50 and std of 15 years.

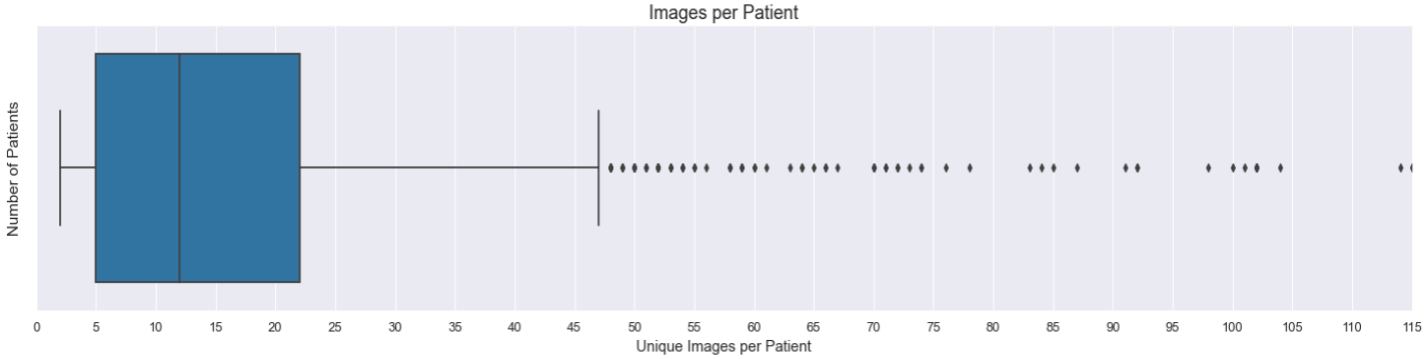

Figure 4: The number of images from each patient varied broadly, showing a strongly right-tailed distribution

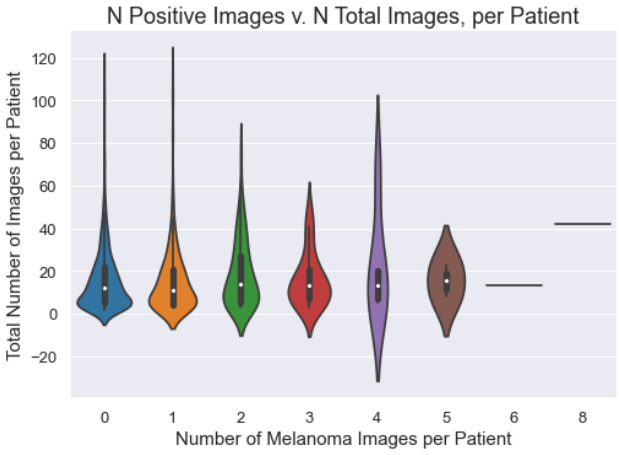

Figure 5: A higher number of total images was not associated with how many of those images were of the malignant class.

Data Handling

To balance the classes for modeling, a subset of the training images was selected through down-sampling the overwhelming negative target images. All images positive for the target class were included (584) as well as an equal-quantity random subset of images which were negative for the target class, giving a total subset of 1168 images which were 50% positive and 50% negative. This training subset was further split in the following way:

1. 20% of the subset training data was held-out as internal validation data. This unseen data allowed us to monitor and visualize model performance on validation data with known labels.
2. The remaining 80% of the training subset data was used for model training. This data was split into 5 folds for cross-validation training of the chosen vision learner, resulting in an ensembled model with makes predictions by finding the average prediction value across 5 trained models.

The images were prepared for training through the use of the ImageDataLoaders datablock class in FastAI, using the "from_df" method. This allowed the metadata stored in the .csv files to be used to specify the labels for each image. Several iterations of batch transformation experiments revealed that simply normalizing the images according to the ImageNet standards resulted in the highest AUROC scores during training. The key word arguments used in the dataloader are as follows:

- df = pandas DataFrame holding the training image data
- path = path where images are stored
- suff = ".png," to specify the file suffixes
- batch size = 8
- device = cuda:0 (in this case, the GPU available in Kaggle Notebooks)
- batch_tfms = Normalize.from_stats(*imagenet_stats)
- label_col = "target"
- valid_col = "fold_{x}_valid"

Model Training

The final training subset of 935 images was used 5 separate times to fine-tune the ResNet34 CNN, each time using a different fifth of the training data for model validation. Each model was fine-tuned over 15 epochs, monitoring for improvements in the AUROC score and reducing the learning rate after 2 epochs without improvement. After all 15 epochs ran, the parameters for the training epoch with the highest AUROC were saved as the best model. This process resulted in 5 separate CNN models for predicting the target class from images, each trained on a slightly different 80% fold of the training data subset. For each validation sample, the final prediction score was taken as the mean prediction value across the ensemble of 5 models.

The mean cross-validation scores for AUROC, error rate, precision, and recall were 0.89, 0.20, 0.79, and 0.81, respectively, as summarized in Table 1, below.

Mean Cross Validation Results n-folds = 5 Class balance = 50:50	
Metric	*Score*
AUROC	0.885
Error Rate	0.201
Precision	0.793
Recall	0.821

Table 1: Mean model metrics for the 5 CNN learners trained during cross validation

Model Validation

Validation of the model was completed along 2 pathways. The primary pathway was through submission of test predictions to Kaggle. In addition, one fifth of the training data was held-out as a validation set in order to validate the model internally and evaluate performance metrics outside of a single AUROC score as returned by Kaggle.

When submitting the ensembled model predictions to Kaggle, the public AUROC score was 0.87, while the private score was 0.85. The similarity between these two measures shows stability in the model performance. We can interpret the private AUROC score to mean that the developed model has an 85% chance of distinguishing between benign and malignant skin lesions.

External (Kaggle) Validation Results Class balance = unknown		
AUROC	*N samples*	*Score*
"Private"	7,687	0.853
"Public"	3,295	0.870
Weighted Average	10,982	0.858

Table 2: AUROC scores for the final model of ensembled CNN learners on the unseen Kaggle competition test set

In addition, the "internal validation set" of 233 unseen images was tested against the model, which allowed us to generate some visualizations of model performance. The receiver operating curve (auc=0.9, Figure 6) and precision-recall curve (auc=0.9, Figure 7) both indicated a skillful learner, at least on this balanced test set. Using a threshold of probability = 0.5, the learner achieved the performance summarized in Table 3, below. Of note, with a precision value of 0.83, the model was able to correctly classify 83% of the malignant images in the internal validation set. The confusion matrix summarizes the classification results (Figure 8).

Internal Validation Results Sample size = 234 Class balance = 50:50	
Metric	*Score*
AUROC	0.9
Average Precision	0.9
Precision	0.83
Recall	0.83
Accuracy	0.83

Table 3: Validation results on unseen data with known labels, "internal validation set"

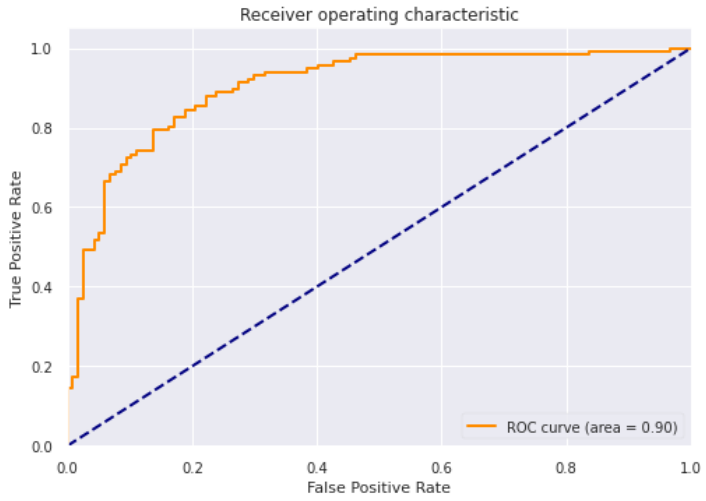

Figure 6: The ROC curve calculated from the internal validation set of 234 unseen images. *The blue line shows the threshold of an unskilled learner, and the ROC curve shows that the ensembled CNN learner is skilled at all classification probability thresholds.*

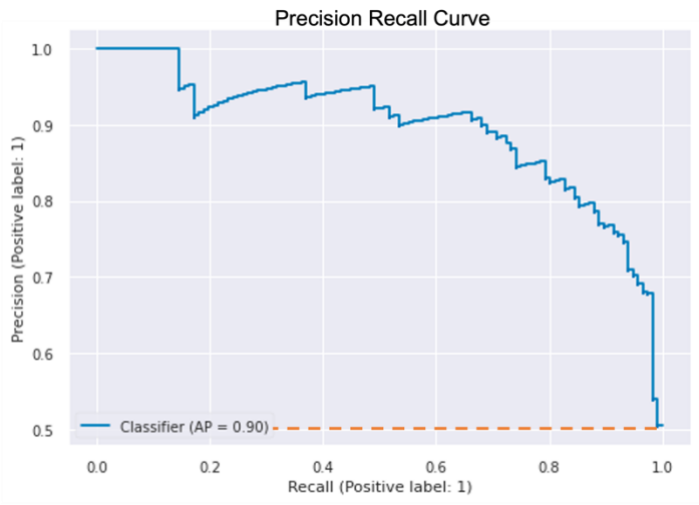

Figure 7: The precision recall curve calculated from the internal validation set of 234 unseen images. The orange line shows the threshold for the 50:50 class balance, and the curve shows that the ensembled CNN learner is skilled at all classification probability thresholds.

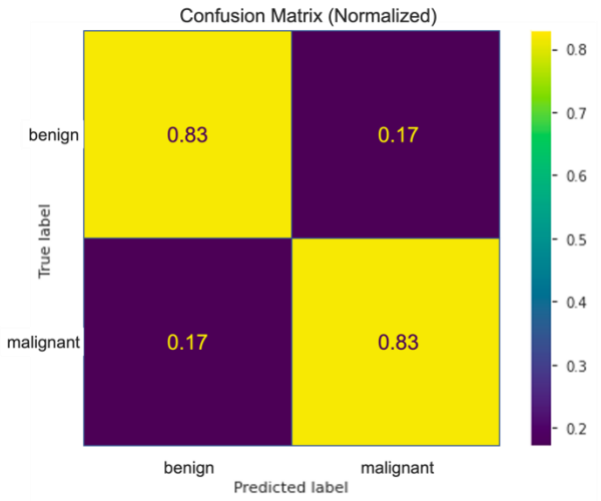

Figure 8: The confusion matrix for 234 unseen internal validation images shows the percentage breakdown for Type I and Type II errors.

Suggested Uses

This model can create value for several main entities: patients, healthcare providers, and health insurance providers.

The first benefit is for the patients themselves; early detection of melanoma greatly improves their prognosis and the severity of the treatments required to resolve their condition. This leads to saved lives and patients who have the agency to skillfully screen for melanoma from the comfort and convenience of home with artificial intelligence technology. When patients are screened and treated early, healthcare providers such as individual doctors and hospital systems experience less system strain. Treatments are less invasive and intensive, and the care needs of the patients therefore are the same. All of this is of great benefit to health insurance providers. The insurance companies are contracted to pay for the care of medically-necessary treatment. It costs less to the insurance company if melanoma is caught and treated early with in-office excision and perhaps some biopsy, versus when it is caught in later stages and full-blown, systemic chemotherapy, radiation and ongoing testing become the needed course of treatment.

My suggested deployment of this model is as a web or mobile application a person can navigate to, upload an image of their mole, and receive a screening result. If their result is positive, the patient would be directed to educational resources about the importance of early screening and treatment and would be directed to make an appointment with a specialist for medical confirmation. This process is summarized in Figure 9, below. A health insurance company could build this out into a specialize app for their members, potentially offering "wellness rewards" for participating in the screening and following up on any recommendations, tracking individual moles over time, and directly linking to providers who are covered in-network for ease of member follow-through. This tool could also be used by primary care physicians, dermatologists, hospital systems, or directly by curious people interested in being proactive about their healthcare.

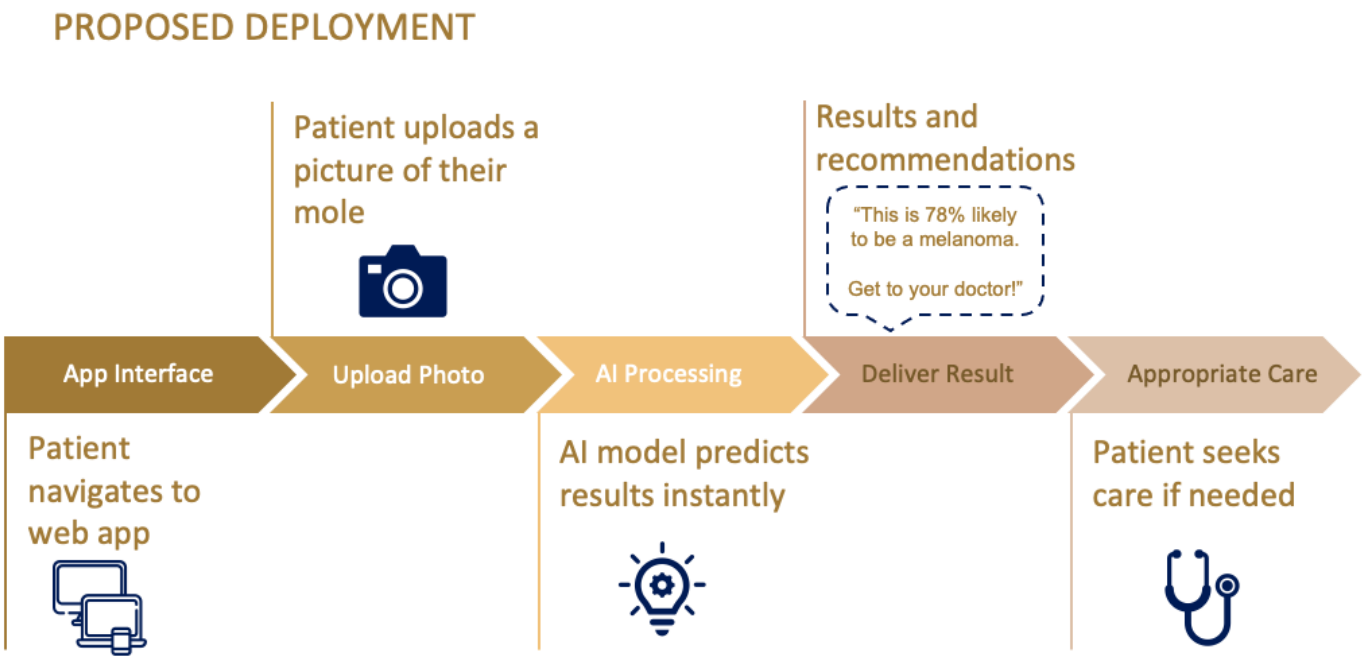

Figure 9: A modest proposal of a deployment pathway to add value for patients, healthcare providers, and health insurance companies.

Limitations and Future Work

One of the major limitations of this dataset and therefore model is the lack of melanin diversity in the patient images. All cases of melanoma are from non-melanated (i.e. “white”) people, as well as all of the observed benign images. While it’s true that the vast majority of melanoma cases are in white people, BIPOC are also afflicted by the disease and often have a worse prognosis due to being diagnosed once the disease is in a more progressed state. Care must be taken to ensure patients with melanated skin do not rely on this tool for medical screening. Unfortunately, medical research involving the use of imaging and light often centers around people with white skin, contributing to racial and ethnic group injustice in medicine. We have to do better by including representative samples in our research and validation!

The current model only takes into account the image data, excluding patient metadata that could strengthen its predictive power. A future iteration could blend image and tabular metadata in the deep learning training process.

In order for this model to become a useful tool, it needs to be deployed as a web app or developed into an app interface. We should keep in mind the experience of a person using the app, and strive to be transparent, informative and interpretable to the lay-person.

A final suggestion would be adding a memory/temporal element to model. Images of the same skin lesion, taken over time, could give valuable insight into any changes occurring and lend data support toward or against a melanoma categorization.

Sources

1. The ISIC 2020 Challenge Dataset https://doi.org/10.34970/2020-ds01 (c) by ISDIS, 2020

   Creative Commons Attribution-Non Commercial 4.0 International License.

   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.

2. *Melanoma skin cancer: Understanding melanoma*. American Cancer Society. (n.d.). Retrieved October 24, 2022, from https://www.cancer.org/cancer/melanoma-skin-cancer.html