Jarom Hatch

Final Report:
Image Classification

Problem Statement

It can often be beneficial to use machine learning to supplement the work done by doctors in diagnosing cases where the problem can be visually seen either externally with images or internally using different X-ray and CT imaging techniques. Machine learning applications by no means replace the diagnosis of a trained medical professional, but can be useful in preliminary screening as well as differentiating conditions, especially those which are often close in appearance, as is the case with this dataset. This can also help to reduce medical staff workload, especially in areas where medical personnel are stretched thin. The data associated with this analysis consisted of a little more than 10,000 unique images of different skin lesions. It also had an associated data table with image id, diagnosis (lesion type), diagnosis method, age and gender of the patient, as well as the location of lesion on the patients.

As part of this analysis, I will be looking at machine learning models geared towards both the image data as well as the image metadata, in order to accurately diagnose certain skin lesions. I'll determine which image classification model is best by the accuracy and loss of the models. For the data table, I'll use the accuracy and precision, recall, and F1 score to judge which model worked the best for the analysis.

Data Wrangling

The data provided was initially spread out between one large file with the 10,015 images and a smaller file containing the tabular data. The tabular data or metadata file contained 10,015 rows, one for each of the images. The columns consisted of lesion id, image id, diagnosis, diagnosis method, age, gender, and location of the lesion.

In order to get the best results from machine learning algorithms, I needed to clean the data. First, there were 57 NaN values in the age column of the metadata. This may have been due to the fact this data was never taken or perhaps the patients wished this data to remain anonymous. Either way, I wanted to fill this data as every row of data would be important in predicting lesion type. To fill this data, I created a new dataset with all the columnar data as one column and determined the amount of unique entries. I compared the rows with the missing ages to the rows which matched this data the best and which were the most common. This gave me the values to impute for the missing ages. When running this process, I also found many rows which contained "unknown" values, which did not appear as NaN values. I used the same methodology to fill these rows as well. This came out to be about 60 rows with unknown values that needed to be filled. I believed this was the best way to impute these values as 60 rows in 10,015 would not be a significant amount and wouldn't skew the data significantly.

After making sure to fill missing values, I also transformed the values to better work with machine learning models. I ended up dropping the lesion id column as I had nothing to correlate this data with and the important information would be retained in the diagnosis type. I then created 7 different variables, 0 through 6 for the diagnosis. I used one hot encoding for the diagnosis method, 0 and 1 for the male/female designation, and again one hot encoding for the locations.

Exploratory Data Analysis

It was important during this analysis to determine what kind of image data I had to work with as well as glean any additional value from the metadata before continuing on to creating machine learning models. I first made sure all the features and targets were converted to numerical values. I then plotted the correlation matrix between all the features as shown in Figure 1.

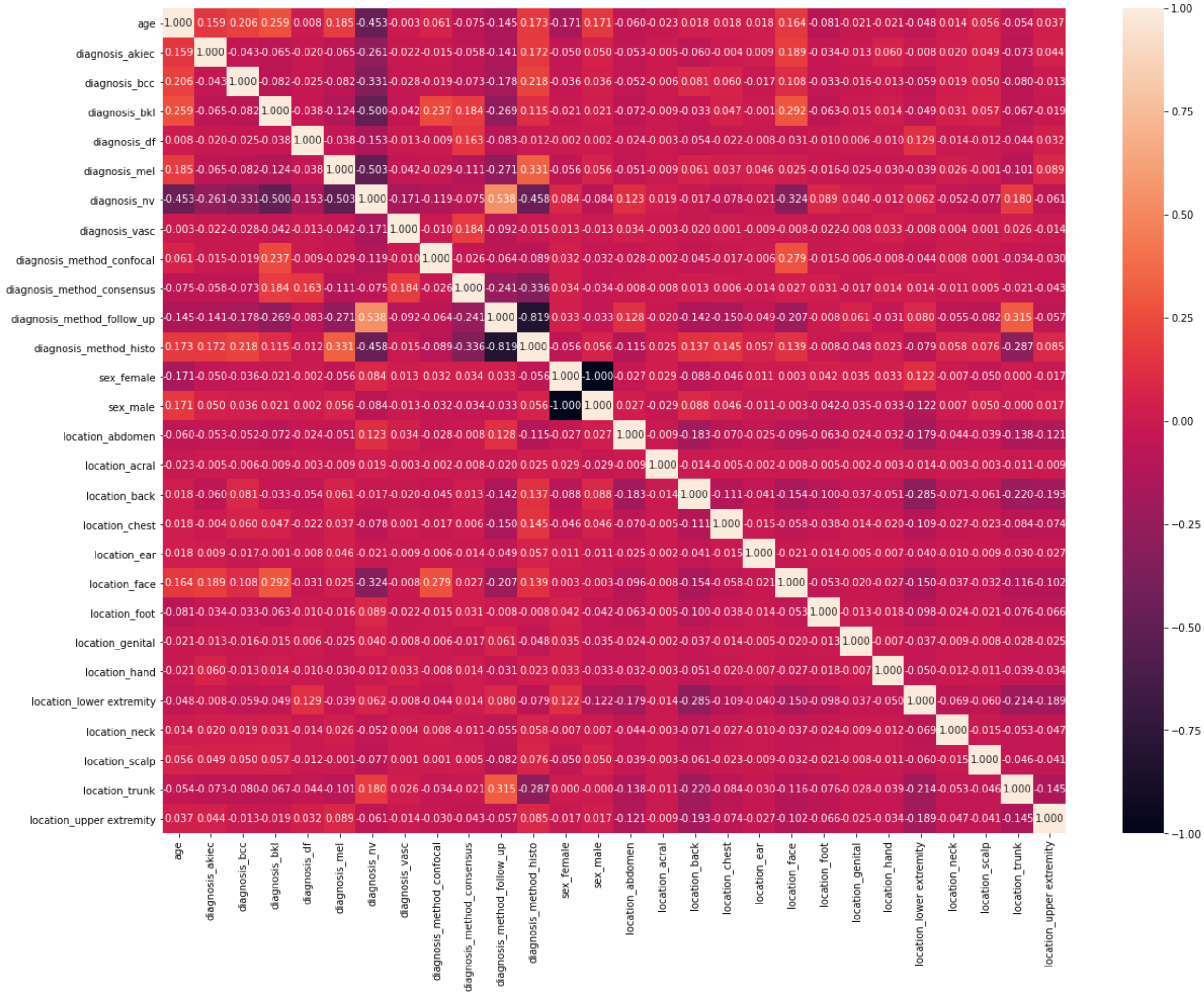

Figure 1: Correlation Matrix

Focusing on the correlation between the diagnoses and the other features, the highest correlations seem to be between the diagnoses and their methods. As can be seen in Figure 1, the nv diagnosis had some high negative correlations with age, and a few of the diagnosis methods. There was also a very high positive correlation between the follow up diagnosis method and the nv diagnosis. I then decided to plot the histograms of the different columns in this dataset. I first looked at the distribution of diagnoses as shown in Figure 2.

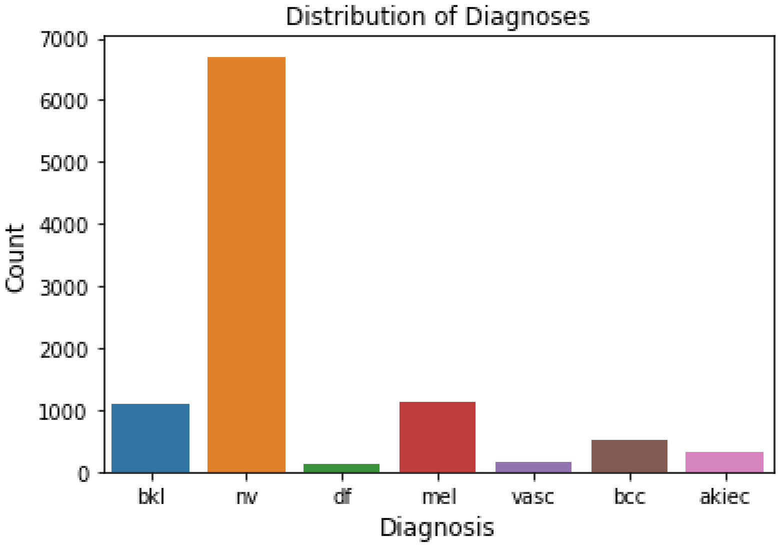

Figure 2: Distribution of diagnoses

Diagnoses consisted of actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus-like keratoses) (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage) (vasc). As can be seen in Figure 2, melanocytic nevi make up the clear majority of lesion types in the dataset. I then plotted the distribution of diagnosis methods provided, shown in Figure 3.

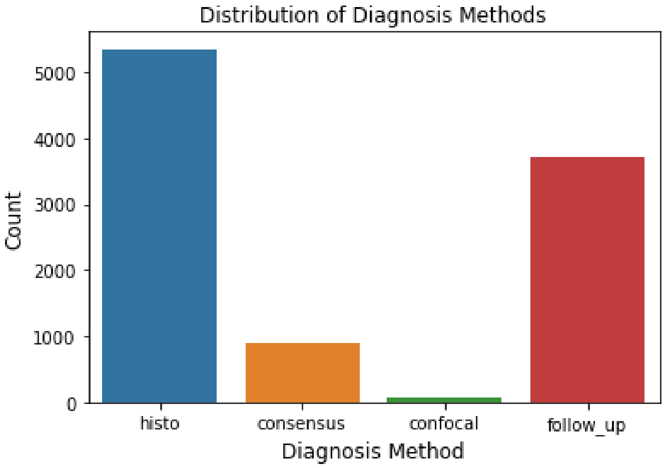

Figure 3: Distribution of diagnosis methods

The different diagnosis methods consisted of histopathology (hist), follow-up examination (follow_up), expert consensus (consensus), and in-vivo confocal microscopy (confocal). The most common types of diagnosis methods were histopathology and follow-up examination.

Next, Figure 4 shows the distribution of lesions locations on the patients in the dataset.

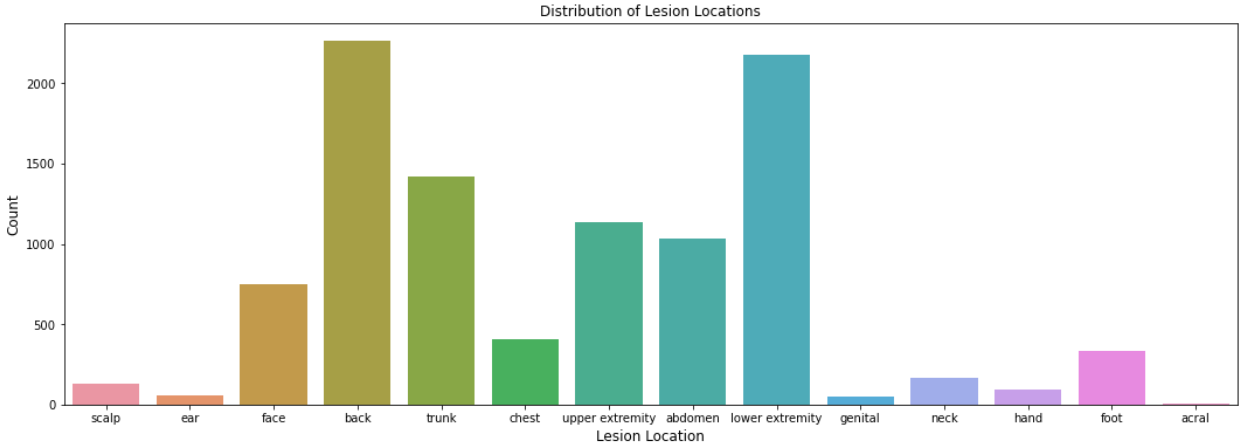

Figure 4: Distribution of lesion locations

Lesion locations were listed as on the scalp, ear, face, back, trunk, chest, upper extremities, abdomen, lower extremities, genital area, neck, hands, feet, and acral areas. The majority of lesions in this dataset were located on about six different locations, with the most being on the back and lower extremities. This data is already very interesting to report on as it could provide doctors with indications of what portions of the body are more susceptible to lesion formation and possibly where people tend to expose the most skin to UV radiation, which may also cause these lesions. Seeing this interesting distribution, we still have the ages and genders of the patients to analyze. Figure 5 shows the gender distribution of the patients in this dataset.

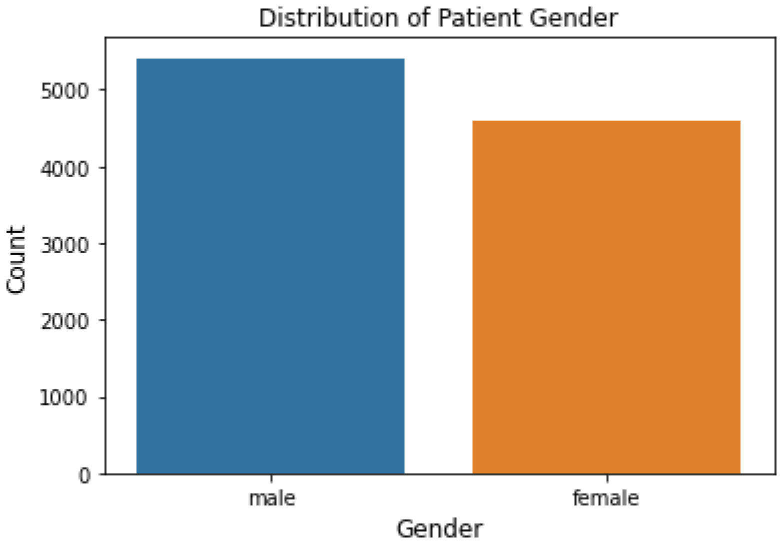

Figure 5: Distribution of patient genders

Figure 5 shows that the data is fairly well distributed between males and females. The lower number of females in this list could indicate that females are less susceptible to lesion formation, they are better at preventing lesion formation, or perhaps that the dataset was just skewed towards men. Further analysis would be needed to determine any significant correlation between lesions and gender. Finally, Figure 6 shows the patients' age distribution.

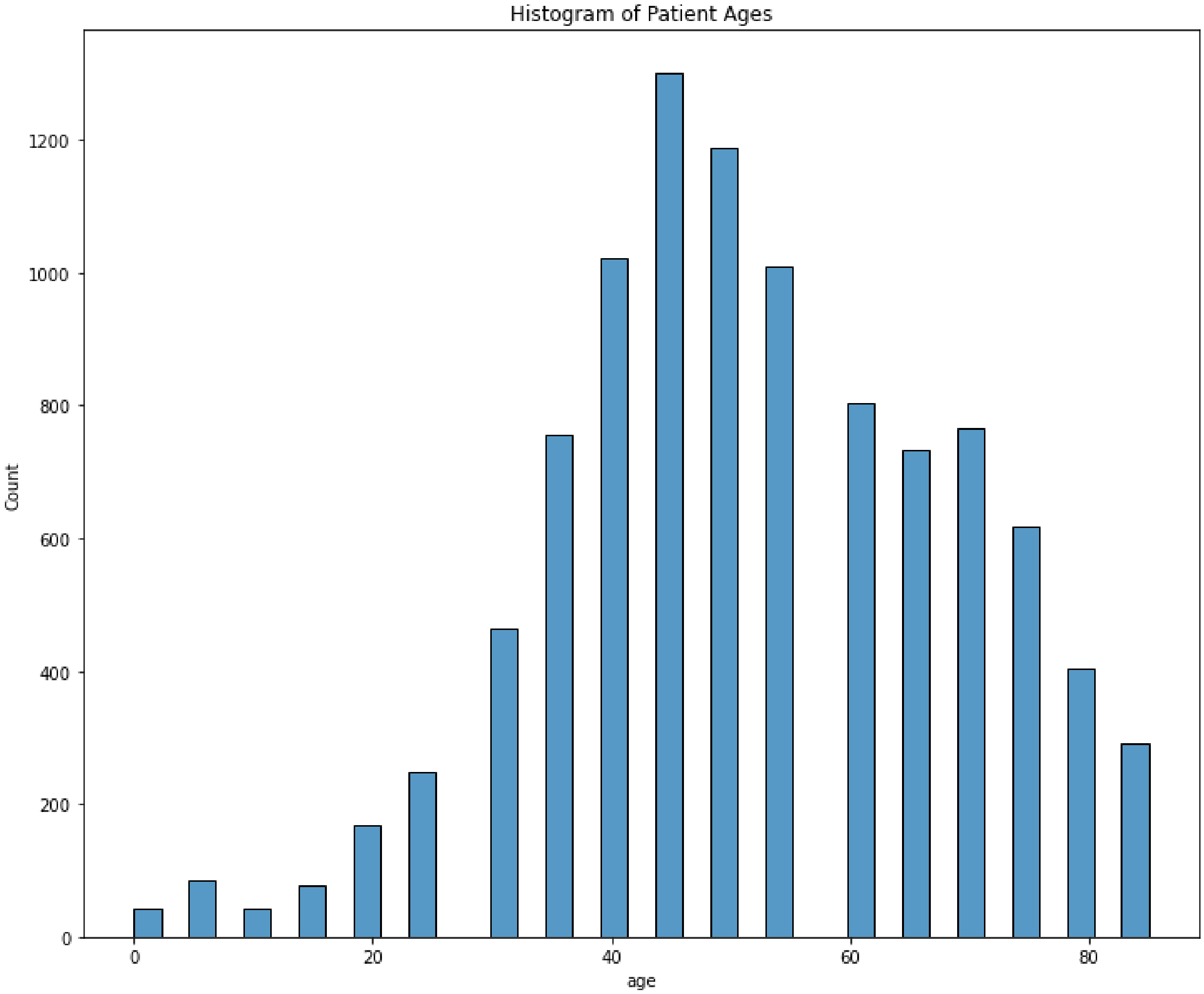

Figure 6: Patient age distribution

The distribution of ages is almost normally distributed, possibly slightly left skewed. The majority of ages for those who have lesions present in this study were in the mid-forties to early fifties. Much like gender, this could indicate some interesting trends between age and lesion formation, but that data would need to be backed up by sampling more patients.

After analyzing the metadata, I also wanted to inspect the images further to determine if they were ready for machine learning and to see if any interesting information could be gleaned from them as well. I first took one image from the dataset as shown in Figure 7. The images in this study were color and 450 by 600 pixels in size.

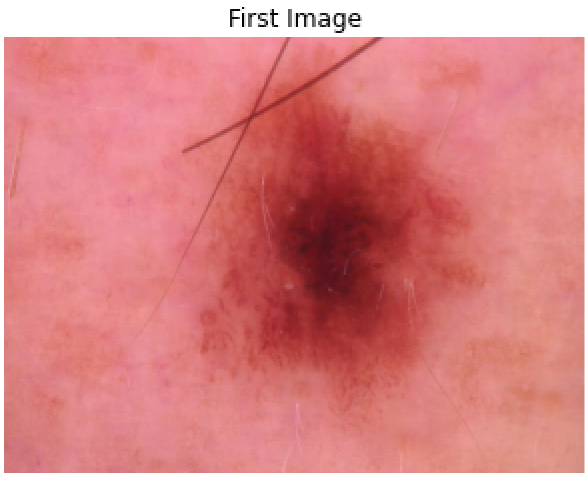

Figure 7: First image in the dataset

I then broke the image into its red, green, and blue color channels to see if any of them would bring out more lesion detail. The results of this are shown in Figure 8.

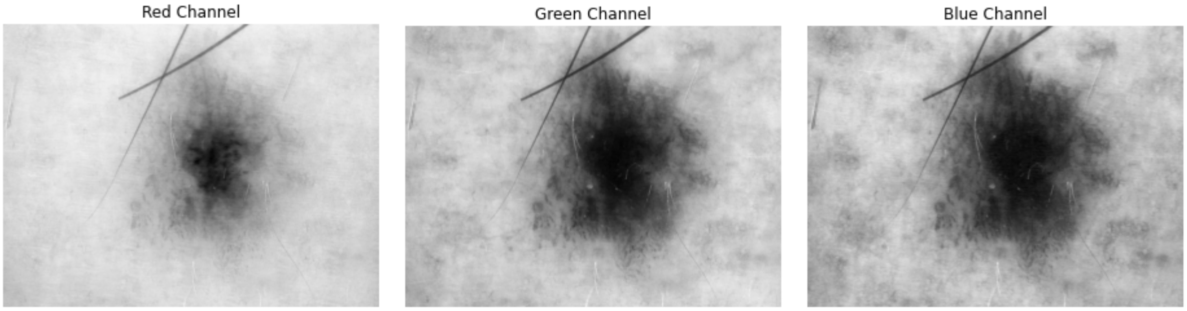

Figure 8: Red, green and blue color channels (greyscale)

It appears that the most data is given by the blue channel, but I wanted to continue looking at other techniques before focusing on this result. The greyscale version of this image, as shown in Figure 9, seems to also show the same amount of detail as the blue channel.

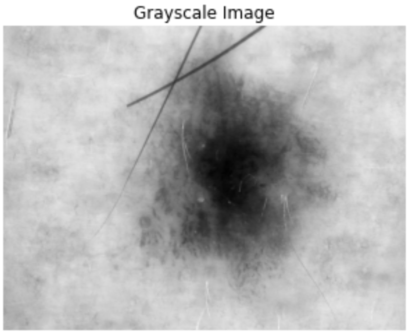

Figure 9: Greyscale image of lesion

I then used different thresholding techniques, as shown in Figures 10 and 11, which brought out some interesting edge and surface detail from the lesion.

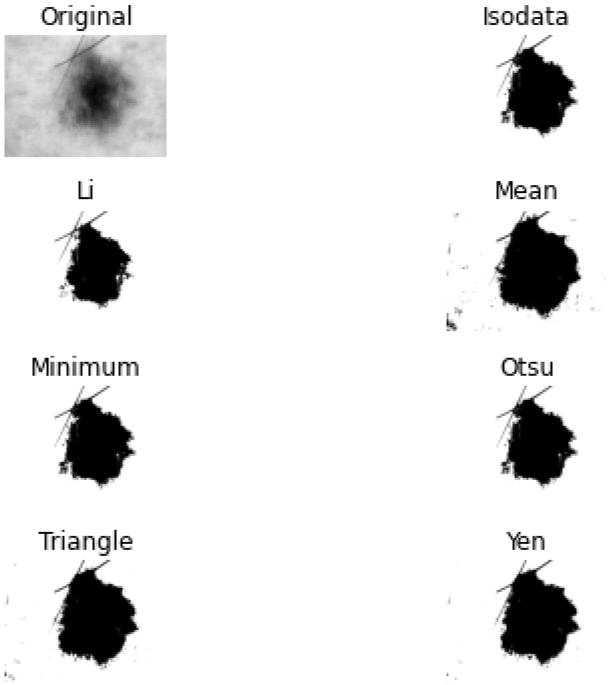

Figure 10: Different thresholding techniques

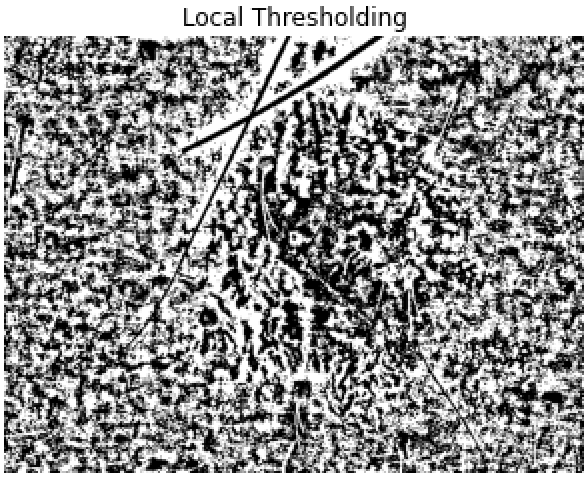

Figure 11: Local thresholding technique

I also found similar results using equalization methods on the image exposure as shown in Figure 12.

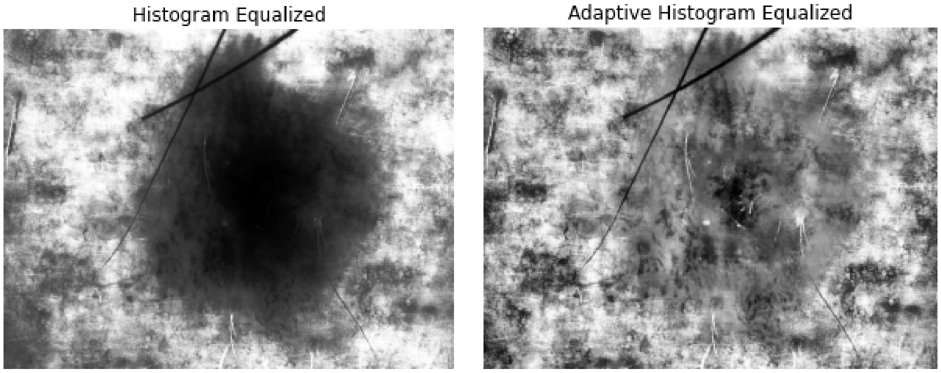

Figure 12: Different exposure equalization techniques

Finally, I also wanted to see if we could really focus in on the edge of the lesion for further processing. I used a counter method and edge detection with Canny to achieve the results in Figure 13 below.

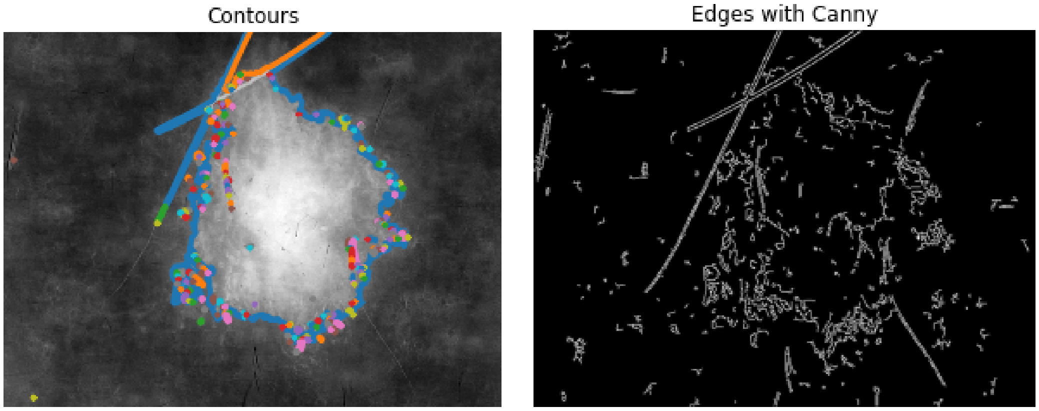

Figure 13: Counter detection and edges with Canny

These methods definitely brought out edge detail as well as underlying surface detail on the skin of the patient. This could definitely be useful for medical professionals in differentiating between different lesion types as well as seeing skin damage not normally visible to the naked eye. However, these methods didn't appear to help the probability of differentiating between lesions in the model, especially since other lesion types also have different colors and similar shapes. Moving forward, I chose to use the original color images for my machine learning models without any modifications.

Modeling

Before modeling, I needed to pre-process both the image and metadata. In order to use the keras TensorFlow model, the images needed to be split into separate folders with each folder being one of the 7 diagnosis methods. Once those were split, I also needed to create placeholders for batch size, image size, and the folder locations. I then used TensorFlow's built-in feature to create a training and testing image dataset. For the metadata, I made sure to convert the diagnosis into numerical values, 0 through 6, for the 7 different classes. In order to optimize my computer's performance, I used an autotune buffer size based on the training and validation images.

I first started modeling the image data. I used a Keras sequential model with 10 layers. In order to further optimize the images, the first layer rescaled the images based on 8-bit pixel color values. I think combined multiple convolutional layers with max pooling layers between them. I finally used a flatten layer and two dense layers to return the image classes from the model. I used an Adam optimizer to compile the model, along with a loss function using sparse categorical cross entropy and the accuracy metric. This initial model used 10 epochs. The accuracy and loss of this model are shown below in Figure 14.

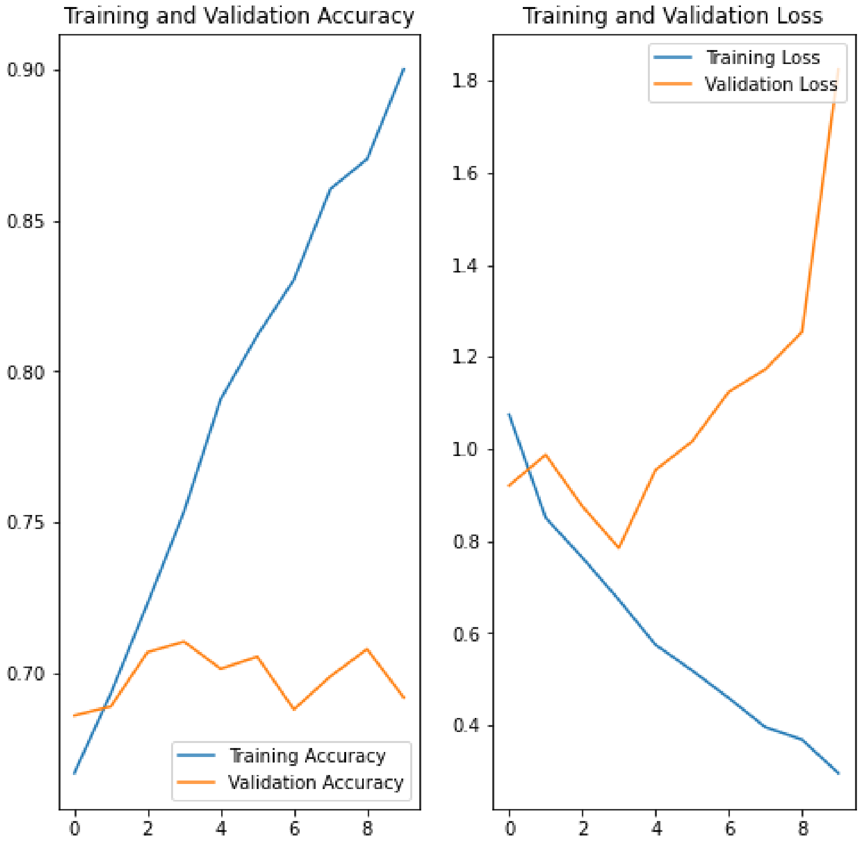

Figure 14: Initial image classification model results

The accuracy of this model seemed to peak after only about 3 epochs around 71% accuracy. The loss visual shows that this model was highly overfit on the data. I believe this was mainly due to the fact of class imbalance and the lack of dropout layer in the model. In order to offset any overfitting occurring, I decided to incorporate some image augmentation by flipping, rotating, and zooming the images at random as well as adding a dropout layer. The next model utilized 14 layers, which included the addition of the augmentation layers and a dropout layer using a value of 0.2. All else was left the same. This time I also increased the number of epochs. The accuracy and loss from this model are shown in Figure 15.

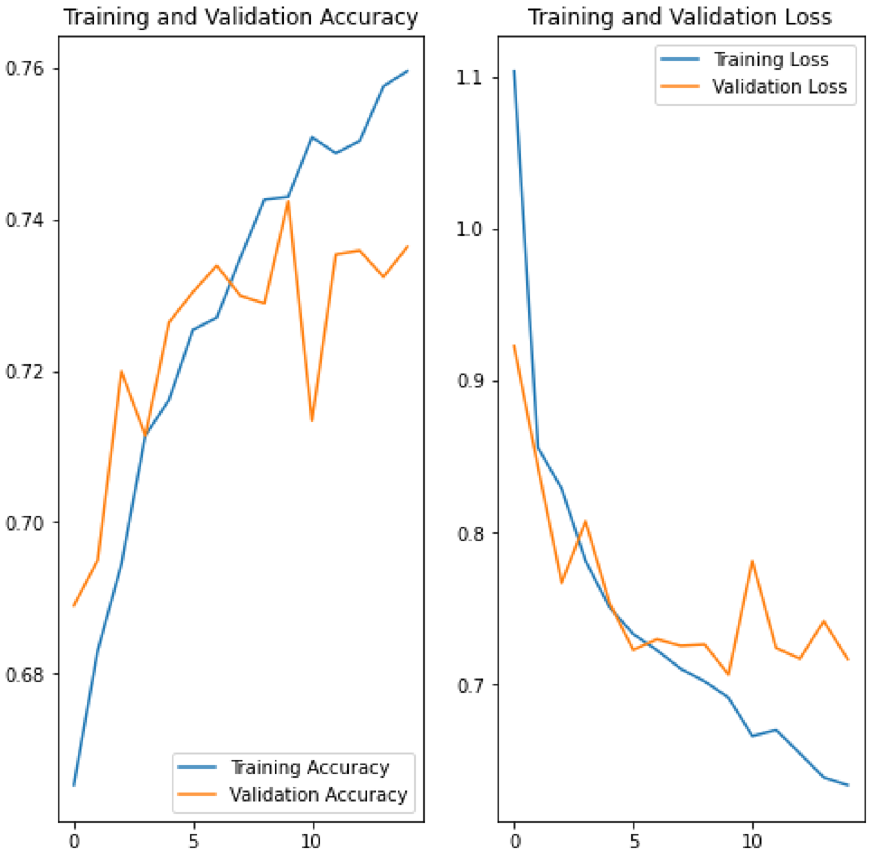

Figure 15: Second image classification model results

This model did much better in terms of loss, and only slightly better in accuracy. Both the accuracy and loss had a large amount of variation depending on the epoch, so I decided to run one more iteration. This last model I tried used the same settings as the second model, but this time a dropout of 0.5. The results from this model are shown in Figure 16.

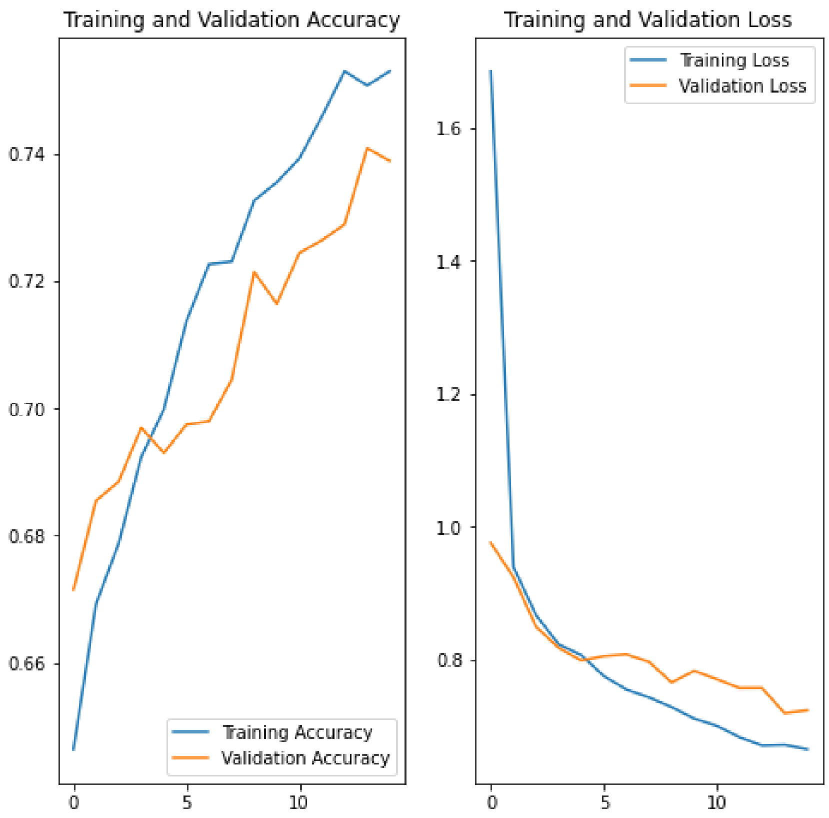

Figure 16: Final image classification model results

This final model had much less variation between epochs and had a low amount of overfitting and the best accuracy, almost 74%. The results from this model proved that it was indeed possible to model this dataset, although the accuracy was not very high. I wanted to compare this result with models based solely on the image metadata to see if I could increase the accuracy further.

For the image metadata, I split the dataset with X being the features and y the targets or lesion classes. I then used the train/test/split method to divide up this dataset with 80% training and 20% testing data. I initially constructed a basic random forest model with only the number of estimators specified at 200. The accuracy from this model came out to about 73%, very similar to the image models. Since there were multiple classes, I also created a classification report to determine how well the model did on each class. There was definitely a noticeable difference in precision, recall, and F1 score between the classes, most likely due to the class imbalance in the data. To try and compensate for this, I decided to also run a random search and grid search cross validation method. Using the best parameters from these methods, the accuracy still didn’t increase and there were only slight changes in the precision, recall, and F1 scores. The resulting metrics from the first basic random forest model are shown below in Table 1.

Target	Precision	Recall	F1 Score	Support
0	0.21	0.12	0.15	69
1	0.29	0.24	0.26	93
2	0.48	0.43	0.45	228
3	0.73	0.29	0.41	28
4	0.36	0.24	0.29	226
5	0.84	0.95	0.89	1338
6	0.46	0.29	0.35	21

Table 1: Metrics from basic random forest model

I also decided to run a few other classification models to determine if any other methods would work better with this dataset. This included an easy ensemble classifier, K nearest neighbor classifier, and a Naive Bayes classifier, all of which had significantly worse accuracy and metrics compared to the random forest model.

After deciding to move forward with the basic random forest model, I also wanted to see which features were the most important in classification. Using the built-in feature importances workflow from the random forest model, I found the following results as shown in Table 2.

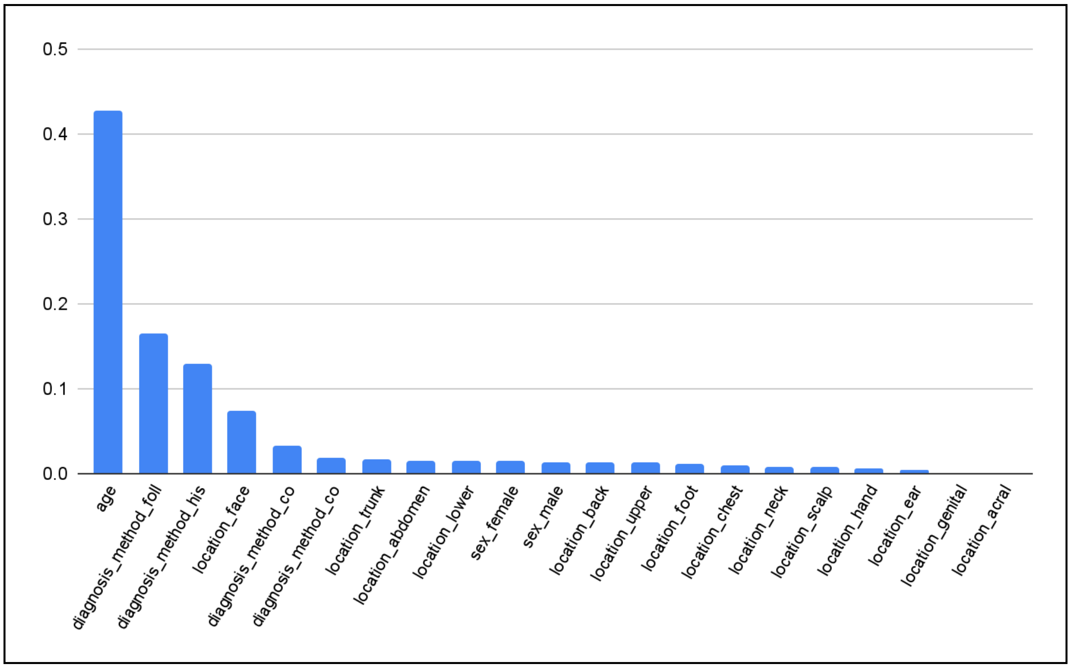

Table 2: Feature importance from random forest model

From Table 2, the top three most important features for this classification model were the patient age, follow-up diagnosis method, and histopathology diagnosis method. The diagnosis methods make sense as they were also the top most diagnosis methods in the dataset. Age would also make sense from a practical standpoint as most patients only get these types of skin lesions from traumatic skin damage, such as from chemical or UV rays, which most younger people are unlikely to have been exposed to. Younger people also heal faster which may help reduce the chance of forming lesions.

Summary

The ability to accurately diagnose different skin lesions based on images or patient information is extremely valuable in the medical world, not only for these lesions but other ailments such as cancers and broken bones. With the power and potential of machine learning becoming more prominent each day, something like this analysis could be condensed into something like a commercial app that could help people catch cancer earlier and know when they need to seek expert medical help.

After applying numerous machine learning models for predicting skin lesion class using both the images and metadata by themselves, I was able to create two classification models. Surprisingly, both models had very similar accuracy of around 73%. This could be due to many factors, most likely the distribution of the data between the different classes. In order to increase the accuracy of these models, perhaps more time and computing power would be needed to search for better hyperparameters. Another improvement for this analysis would be to increase the number of samples, especially for those classes which are highly imbalanced. I could probably have reduced the number of samples in this analysis until the classes were more even, but then I would lose a large amount of important classification data, especially when it comes to the details provided by the images. Perhaps another issue was the close visual similarity between the different types of lesions. More analysis would be needed to see if more distinguishing details could be brought out of the images or perhaps clarified by medical professionals. Regardless of these results, this project demonstrated my ability to work with diverse types of data, clean and prepare that data, and finally build machine learning models tailored to each data type.