Skin cancer detection

Problem statement:

Skin cancer is one of the most common kind of cancer in US. One in five Americans may develop skin cancer in their life, and currently over one million people are living with the disease. Despite such high number of patients, the mortality rate of skin cancer is very low as long as it is treated before the cancer spreads, with the five-year survival rate of 99%. However, five-year survival rates drop sharply after it evolves to regional and distant stage, to 64% and 23% respectively. Therefore skin cancer must be detected and treated early.

Early stage of skin cancer can easily be confused with harmless mole and therefore can be overlooked. With the increased number of potential patients, there needs to be a more accessible way for people to diagnose themselves to find out whether they are at risk of developing malignant skin cancer.

It is well known that malignant skin cancer shows specific characteristic that may imply its danger level called ABCDE rule. Asymmetry, Border, Color, Diameter, Evolution. By utilizing these attributes with image classification, I can build an algorithm for people to check if the mole growing on their arm is cancerous. While not a perfect way to diagnose the disease, it helps to let the potential patient know if they should pay a visit to the doctor.

Dataset:

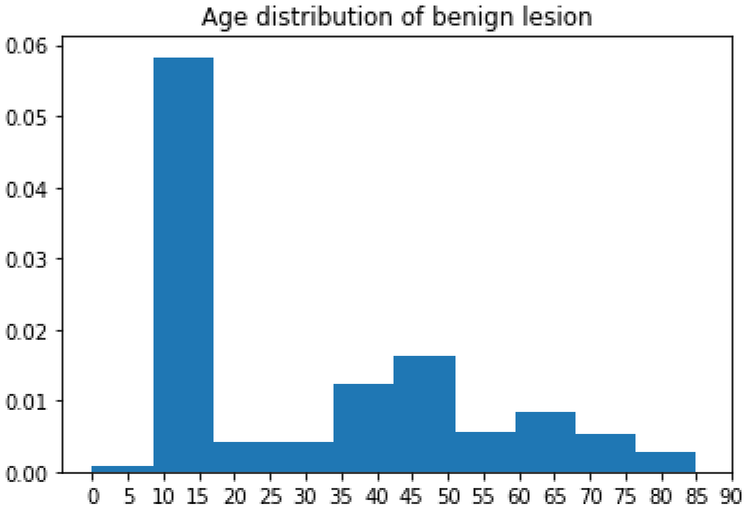
To build this algorithm, I used image data from International Skin Imaging Collaboration (ISIC) where individuals from both academia and industry provide digital images of benign and malignant melanoma to educate other professionals in recognizing the lesions.

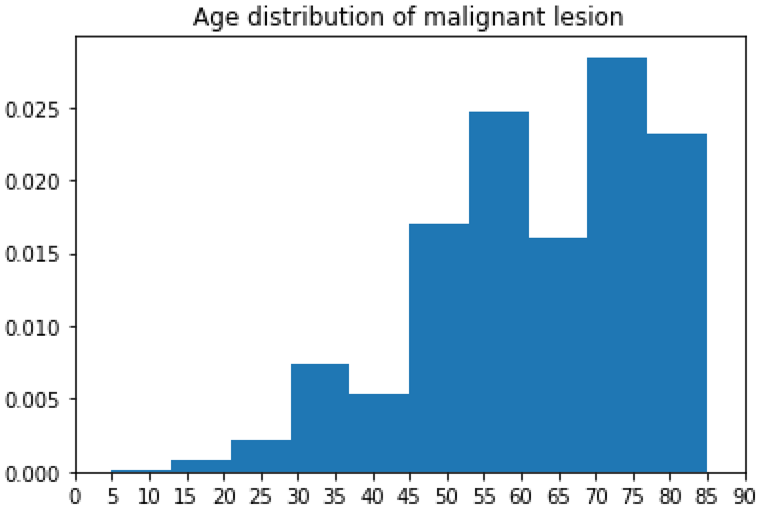
The data was collected via download link provided by the ISIC website (https://www.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery). This includes meta data where it describes whether or not the lesion in the image was determined to be malignant, size and location of the lesion, age and sex of the patient, and the name of the institution from which the image was taken. Such data was used to know more about the general trend of the disease such as where the lesions are likely to appear and at what age range the lesion is likely to be malignant.

In order to perform the image classification, the image will be filtered via gaussian filter, normalized, and in case of color analysis, be divided into red, green, and blue images.

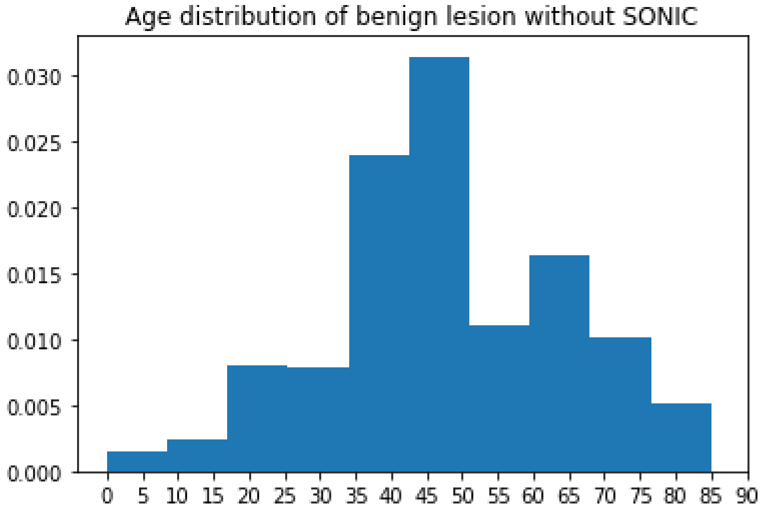
Exploratory analysis:

For initial exploratory analysis, metadata was mainly used to get the idea for skin cancer. The column on the metadata that mentions whether or not the lesion is malignant contains 7 items as sometimes the lesion could not be identified and was marked ‘indeterminate’. For the sake of the analysis, only data marked either “malignant” and “benign” were used.





Regarding age, it is clearly shown that the younger patients are much more likely to have benign lesion compared to older counterparts. There are a significant number of teenage patients with benign lesion, which is due to the dataset “SONIC” which consists entirely of benign lesion from children.

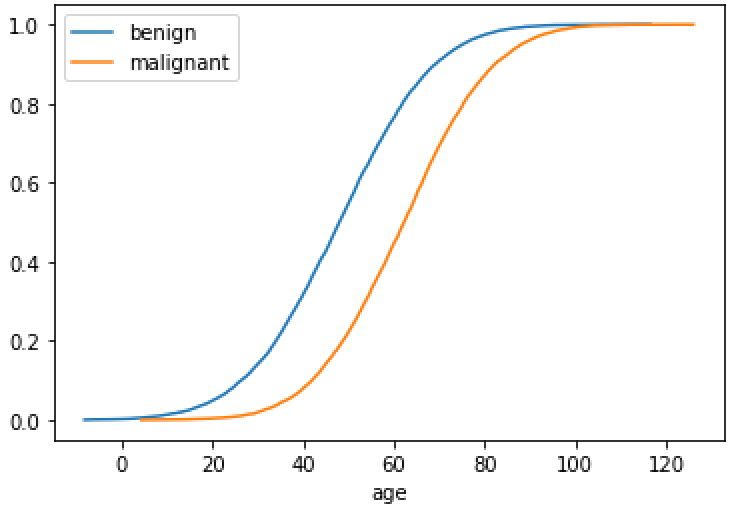


Because SONIC is a set of data with only benign lesion from specific age range, they will be removed for more general data exploration.

Mean and standard deviation from benign and malignant group are as follows:

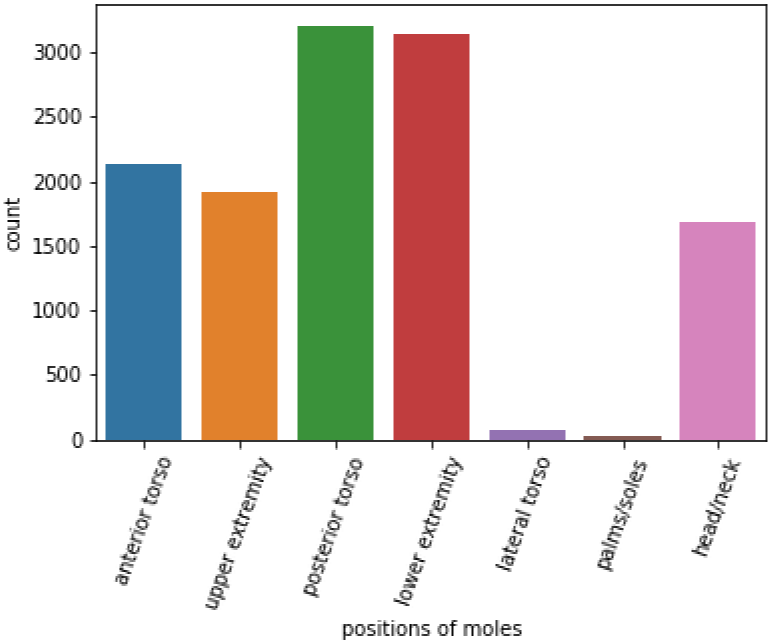
	mean	Standard deviation
Benign (without SONIC)	47.85	16.58
Malignant:	62.00	15.78

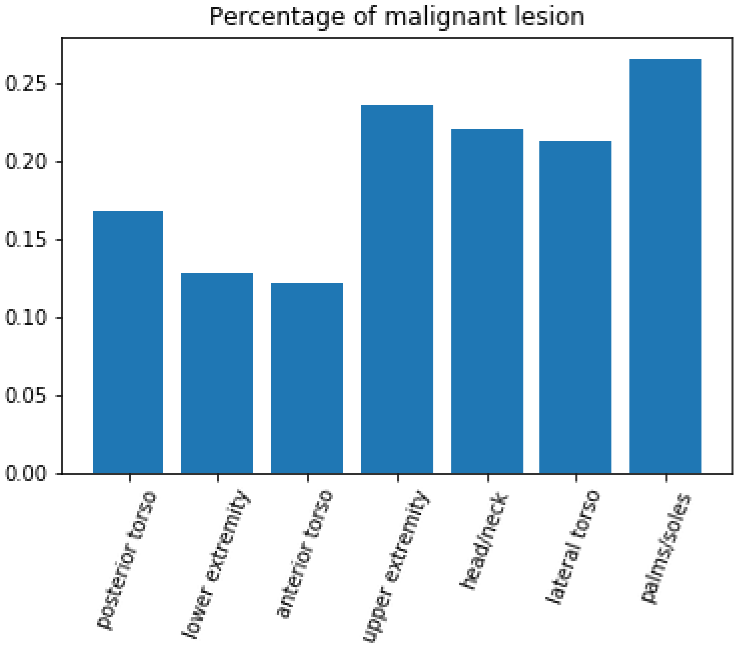
Therefore I can conclude that older patients have a higher chance of having malignant lesions. I can further support this via ecdf graph and performing independent t-test



I can see that benign graph leans more towards younger while the malignant graph is leaning more on the older side. Also the result of the t-test shows that the difference in the mean between the two data is different enough to be significant ($p<0.05$). Therefore I can safely conclude that older patients are more in danger of getting skin cancer.

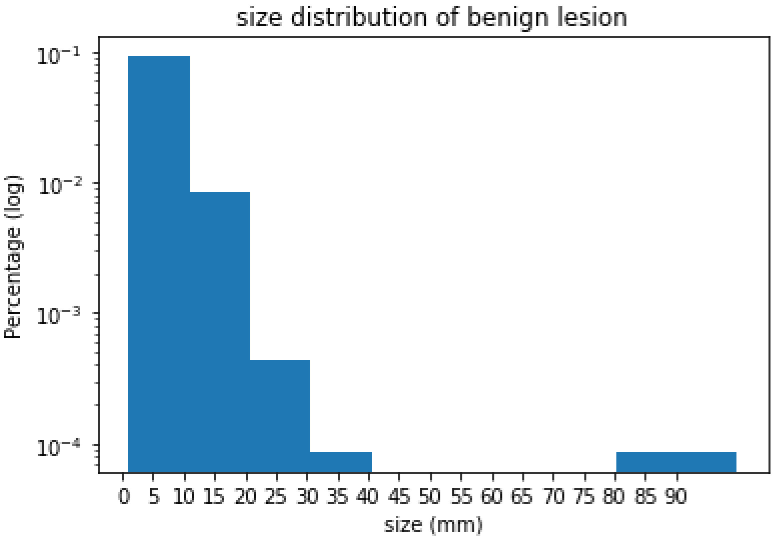
Regarding the position of the lesions, the most common place a person finds a mole is posterior torso followed closely by the lower extremity and then the anterior torso. However when calculating how likely those lesion are to turn out to be malignant, the lesion from palms or soles

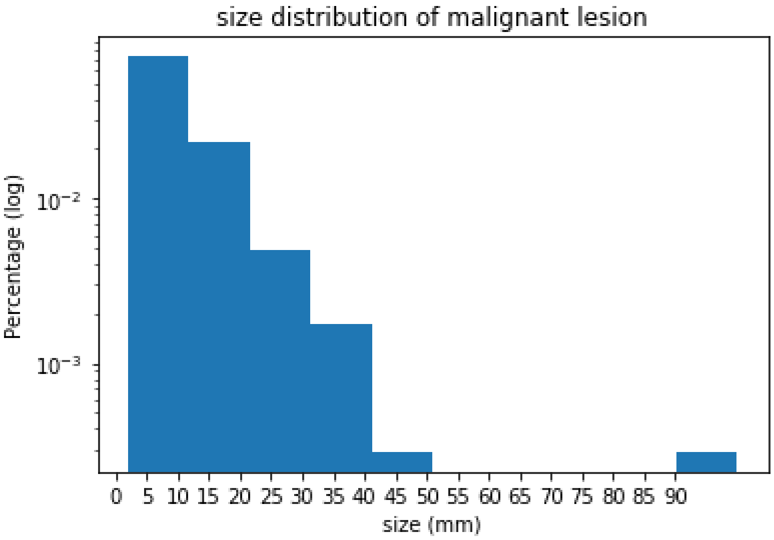


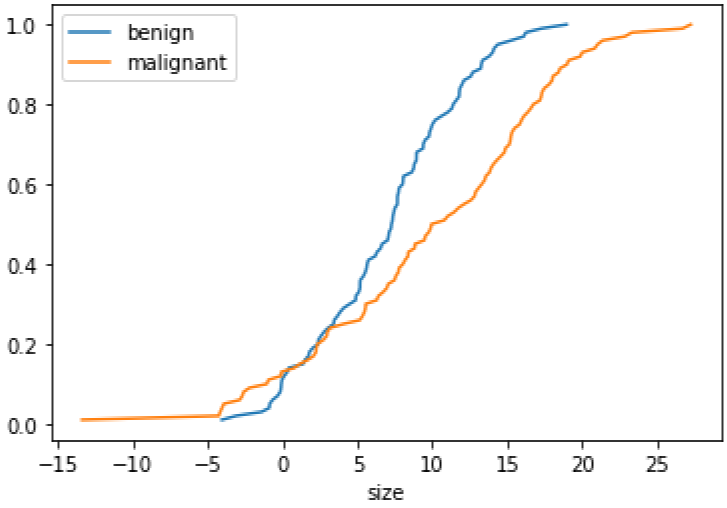


comes out on top, followed by upper extremity and head or neck. This might be due to melanoma developing from exposure to UV light as palm, head, neck, and upper part of the body is more likely to be exposed to direct sunlight.

Size of a lesion is one of the significant factors in determining whether or not a lesion is malignant. Unfortunately, only 1500 images had a metadata that included that information. However, it still showed signs that the diameter is indeed an important factor in determining whether or not the lesion was malignant.


size distribution of benign lesion


size distribution of malignant lesion



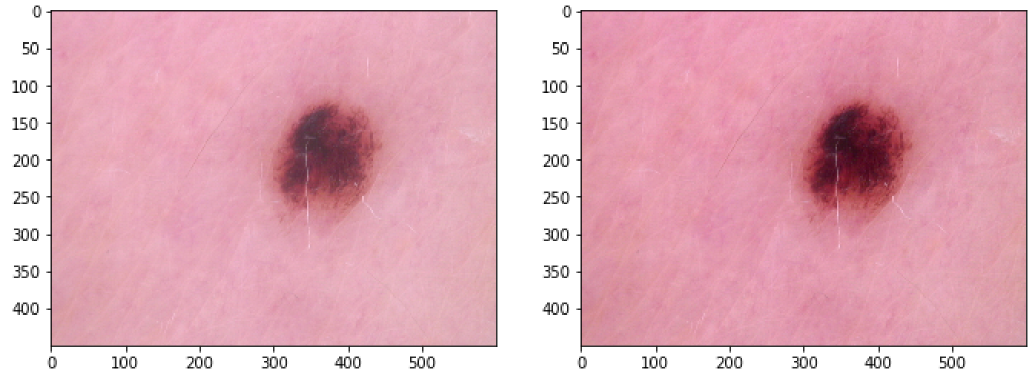
As I did with age of patients, ecdf and t-test was performed on the data, and was shown to agree with our conclusion.

Image analysis was performed on the images to find out the effect symmetry, border, and color of the images has on determining the malignant lesions. Color was divided into red, blue, and green.

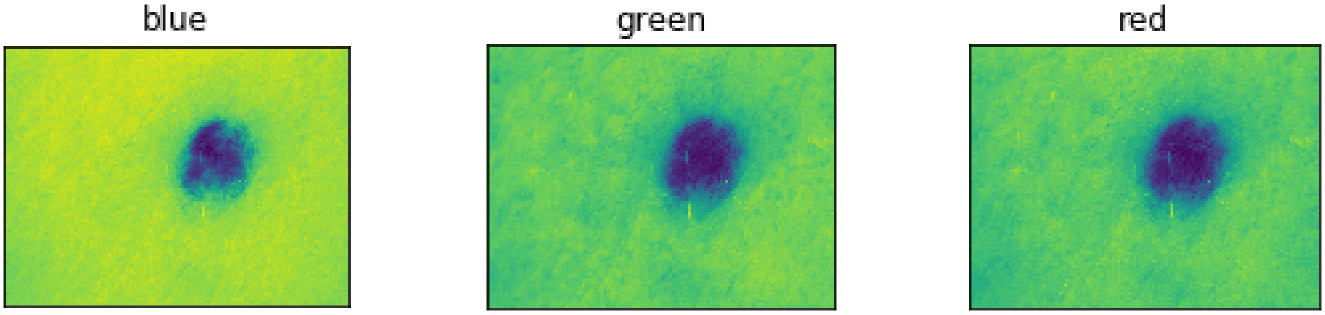
The general preprocessing of the images consists of enhancing the contrast of the image, filtering out the noise, and thresholding the image to segment the lesion from the background skin.

Enhancing contrast:

Normalization of images allows the contrast in the image to become more sharp by making the dark parts more dark and light part more light. This helps with thresholding the image segmentation.



Then the normalized image has the color channel split into red, green, and blue channel. Because human skin naturally have more reddish hue, blue channel usually gives better contrast.

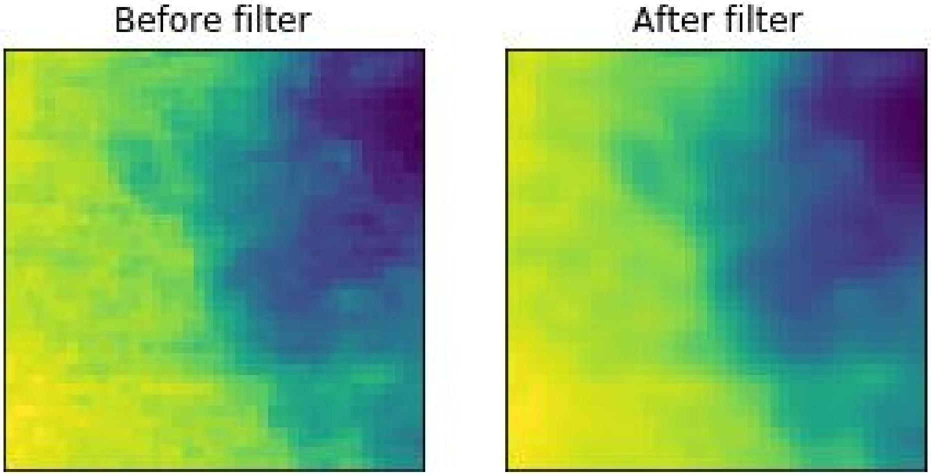


Filtering for noise reduction:

Denoising an image helps to enhance image structure and get better results for edge detection which is necessary for image segmentation. Gaussian filter is often used for this purpose where gaussian function is applied to each pixel of the image.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Formula used for gaussian blur



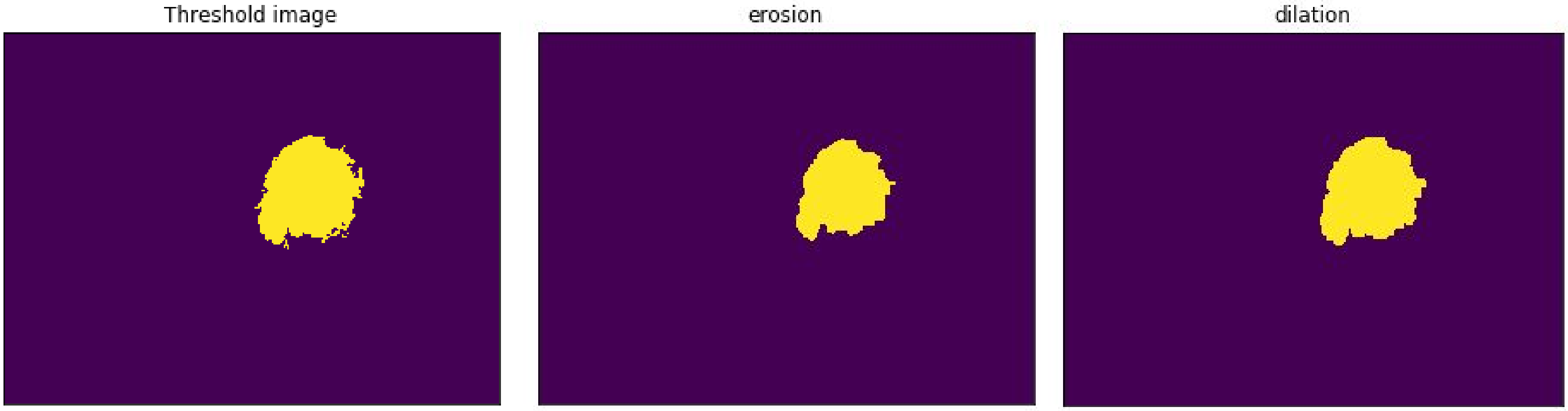
From the image we can see that the image became more uniform with no speck of other color mixed in.

Thresholding for image segmentation:

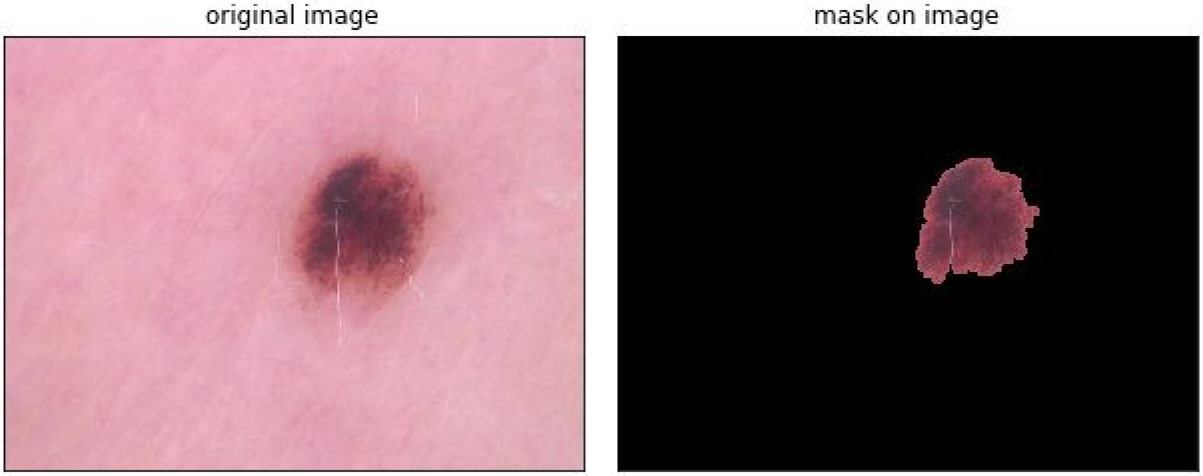
Thresholding takes the grayscale image and make it into a binary image, i.e., if a pixel have intensity that is less than a fixed constant, it will be classified as black pixels. The simplest method will be for the user to set the threshold intensity, but that will deter us from automating the process. For our purpose, Otsu's method of automated thresholding was combined with binary method as it gave the most accurate results.

Otsu's method divides the image intensity range to foreground and background and picks the threshold between the two peaks whereas binary simply takes the intensity that is higher than half the maximum intensity and set it as white. By combining the two, I get the best result that covers most of the lesion.

In order to smooth out the edges of the lesion, the segmented parts are eroded and then dilated. This helps to make the segmented parts detach from the nearby noise that might have appeared due to having similar light intensity.

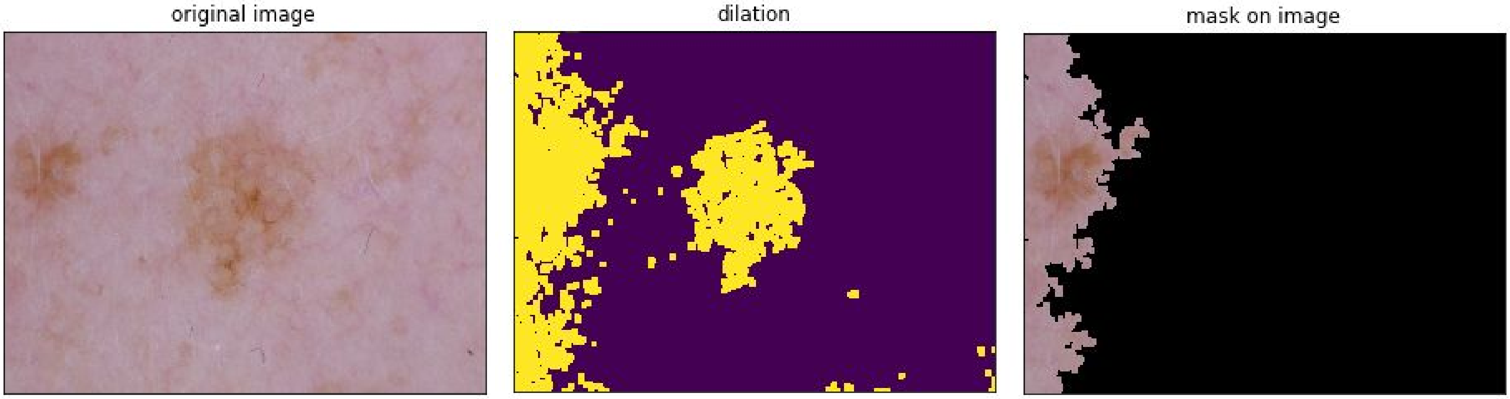


This binary image will then be used as a mask to identify the exact region of interest for cancer analysis.

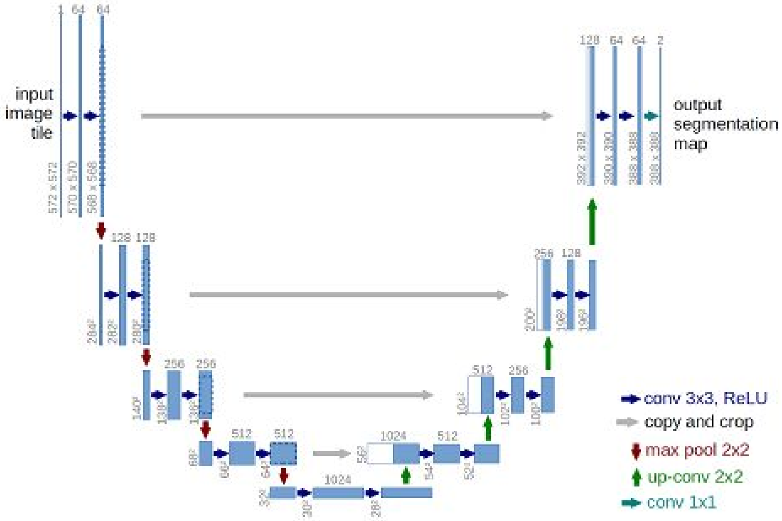


Problem with this method over multiple images:

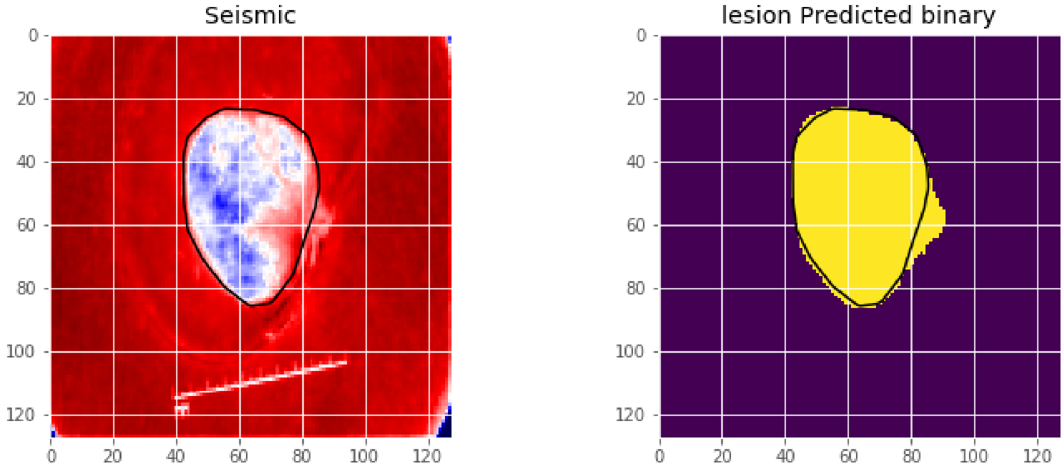
This method has an issue with lesions that are less defined and whose background has different lighting.



Some of the images like this were screened to discard the contour that has long horizontal or vertical edge that is straight for more than 60% of the total length of the image. This helped filter the contour that was stuck on the edge, however as there were still problems on the images that could not be weeded out by machine alone, I tried a different tactic and used a UNET algorithm to segment the images.

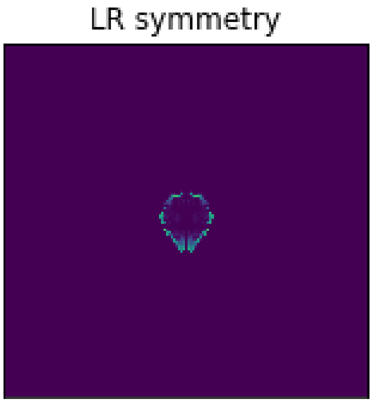


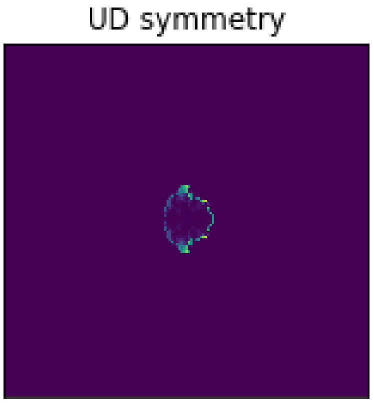
UNET model for image segmentation was developed for biomedical application therefore this was optimal for my current project. When applied to a lesion image, it identifies where the lesion should be and makes a mask to apply to the image.

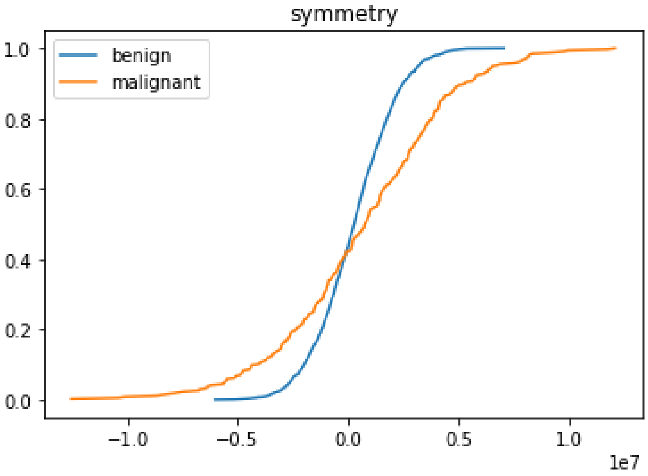


After getting the segmented images, image analysis on symmetry, border, and three different colors was performed.

Symmetry is measured by turning the lesion on its major axis and checking how the lesion differ when flipped on horizontal or vertical line.

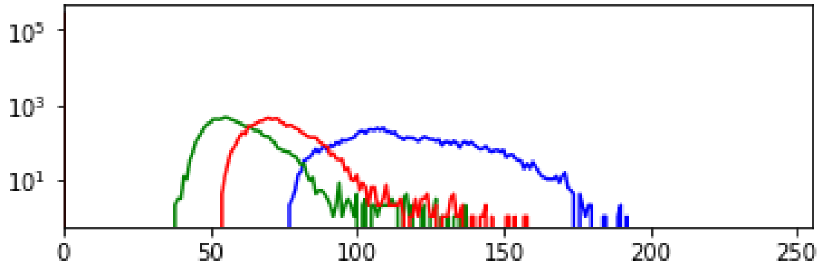


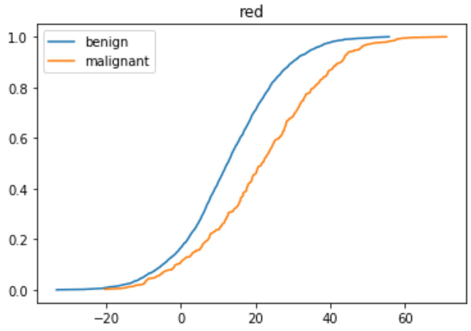


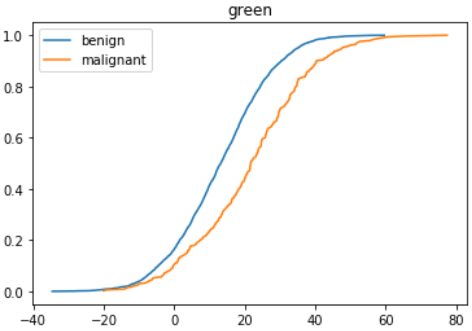


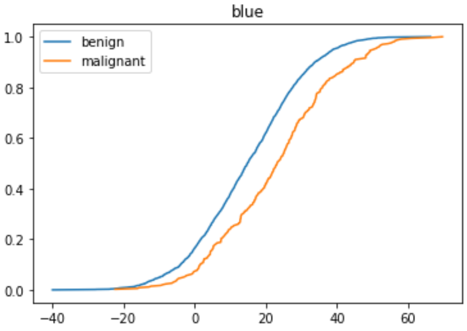
The more difference the lesion has to its flipped part, i.e., the more asymmetrical it was, the more likely it was to be malignant.

The three colors were measured by the standard deviation of each spectrum. The higher the number, the more diverse the color of the lesion which implies that it is malignant.

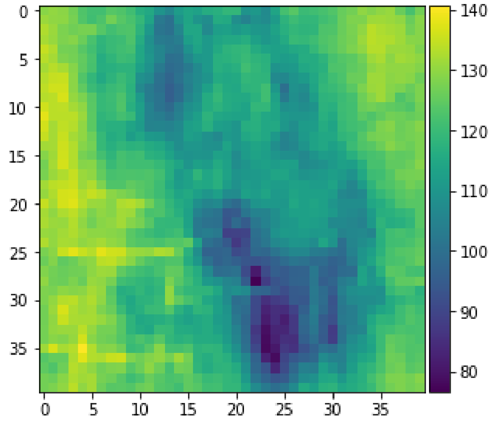


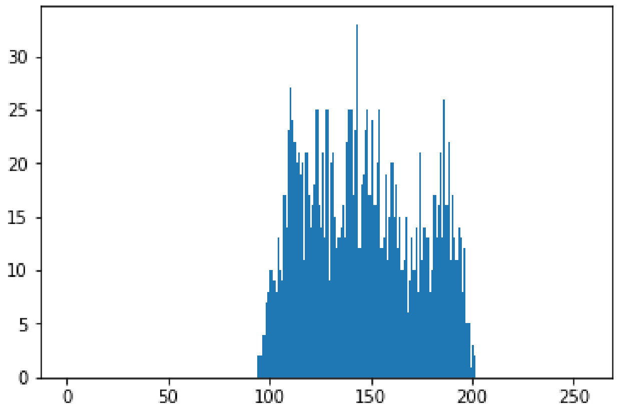


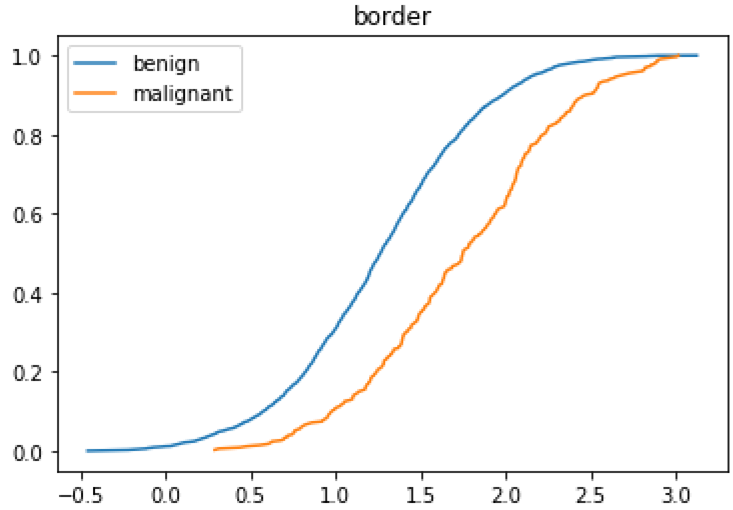




In case of border, I made a small window to go around the contour of the lesion and calculate the mean of the gradient inside that window.



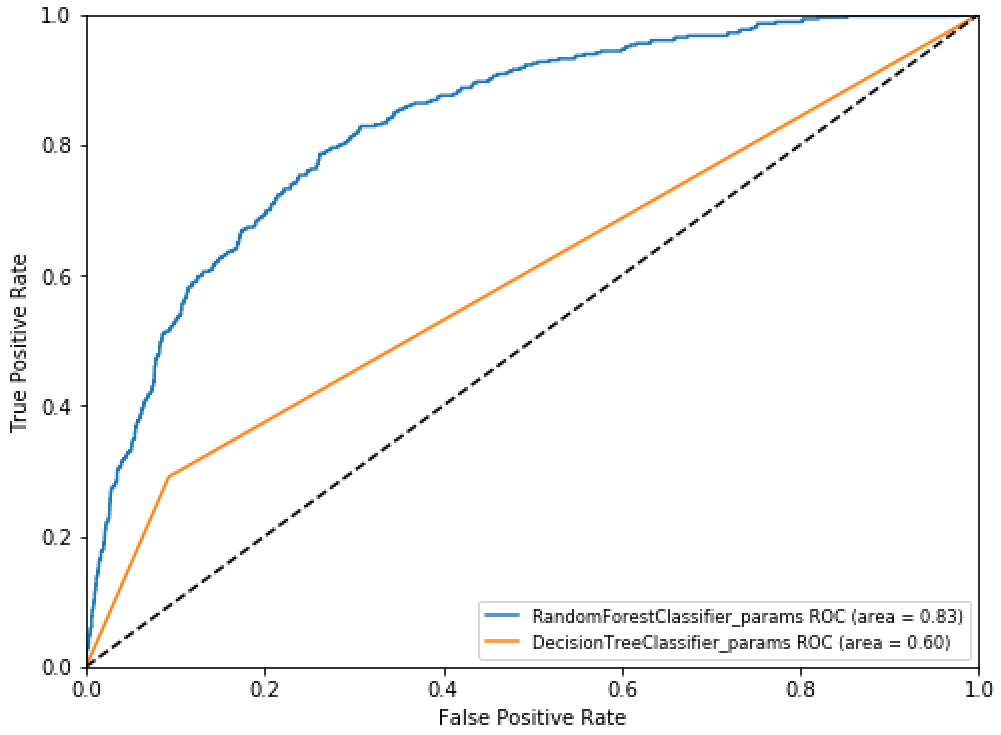




The more gradient it has, the more likely it is to be malignant.

All the features show a significant difference between benign and malignant ($p<0.01$).

With these information, I performed decision tree classifier and random forest classifier to identify the malignant to benign lesions. After training the model, it was calculated for AUC score, and it was plain that random forest classifier is working much better than decision tree classifier.



However, in the case of border it took a very long time as there were over 20,000 images and for every image to have the border gradient be calculated took at least 30 seconds per image which roughly translates to 167 hours for just the border identification. Therefore, I tried a different tactic of utilizing transfer learning.

Transfer learning for image classification

Transfer learning is using a pre-trained model and only training the last few layers for the data I have. Because I am classifying the images with deep learning, the features are generated by the algorithm and it does not require much human input. Also because it only trains the last few layers, the time required for the training is shortened by a significant amount (around 30 min).

For transfer training, 2000 images were used from each benign and malignant lesions. The accuracy and the cross entropy was provided during the training process. I performed the training on “regular”, i.e., images that has not been altered and their size can range from 270000 pixels to , “resized”images that were resized to 128x128, and “regular+resized” which consists of 1000 images from regular and resized each. This was to check if adding resized images will change the result in any way.

	Train accuracy	Validation accuracy	Cross entropy
Regular images	92	77	0.1921
Resized images	87	79	0.3245
Regular+resized	86	83	0.2743

The cross entropy of the regular images was the lowest among the three instances although the validation accuracy is shown to be increased in “regular+resized”. However, since the validation accuracy and train accuracy tend to fluctuate between 77 to 90 when attempted several times, it may be safer to assume that just having regular image to train the model is better. This is probably due to the fact that regular images have more diverse range of image sizes and can be trained to identify the lesion even in different sized images. Having images of same size takes that diversity away and can lead to the decrease in accuracy.

Compared to the accuracy from random forest classifier (83%), the accuracy does not increase by a lot but the time it took for the training is decreased by a significant amount and therefore it is worth changing to this method.

Conclusion

In this capstone project, I was able to look at how skin lesion may be classified to be identified for malignancy. More traditional methods were used at first, however due to the limitation of not having enough identifiable features that can differentiate malignant from benign lesions, deep learning algorithms were incorporated in image segmentation and due to the training time required, image classification was performed with deep learning algorithm as well.

Through this experience I was able to recognize the advantage of developing a deep learning algorithm for initial screening of malignant skin cancer and how to utilize transfer learning to lessen the training time required for such problems.