# Capstone 2

Zoe Shim

# Introduction

- Skin cancer is one of the most common cancer in US
- Early detection => 99% five-year survival, distant stage = > 23%
- ∴ Faster more accessible melanoma detection is needed

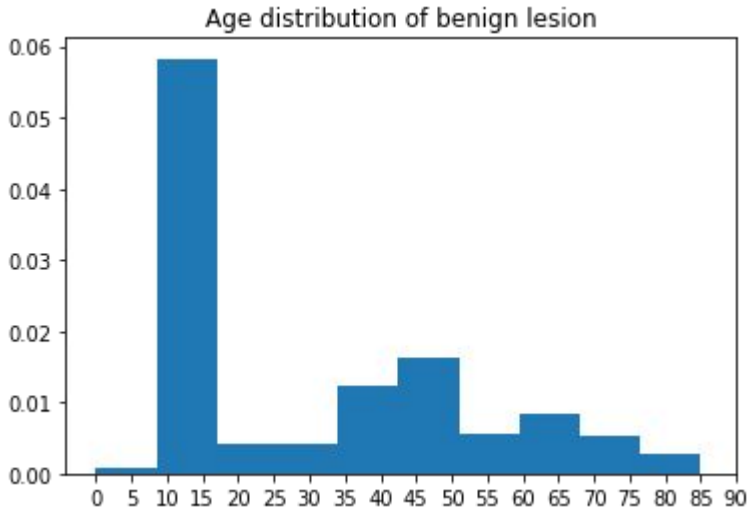
# Data wrangling

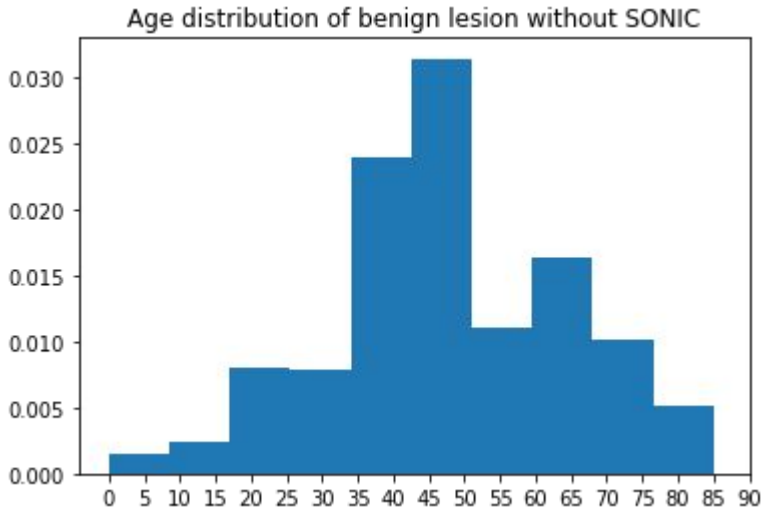
- Download image data and their metadata from International Skin Imaging Collaboration
- Load metadata to dataframe form for EDA
- Load images and normalize
- Split the images' R,G,B channels for color analysis
- Apply gaussian filter and segment lesion using the threshold acquired via the filter
- Remove images that does not have enough contour to analyze

# Data Exploratory Data Analysis

- Size of the lesion
  - Bigger the lesions, more likely it is to be malignant (D in ABCDE)
- Age of the patients
  - Removed certain dataset due to bias introduced
  - Older the patients, more likely to have malignant lesion
- Location of the lesion
  -

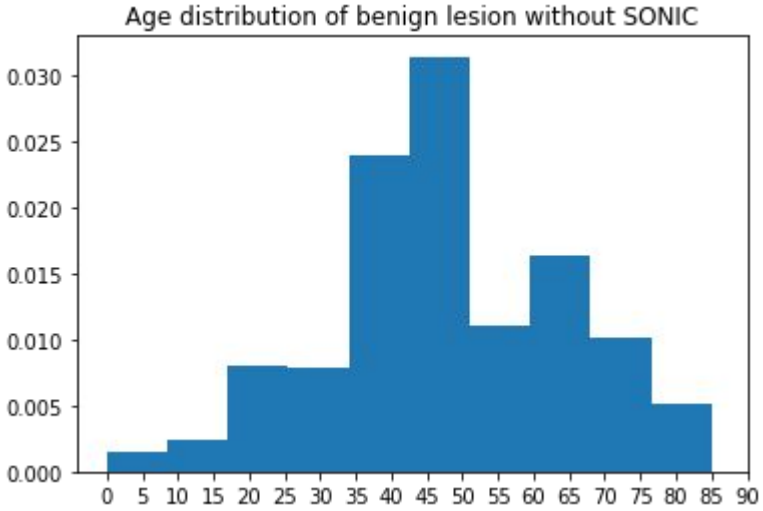
# Benign vs. Malignant: Age distribution

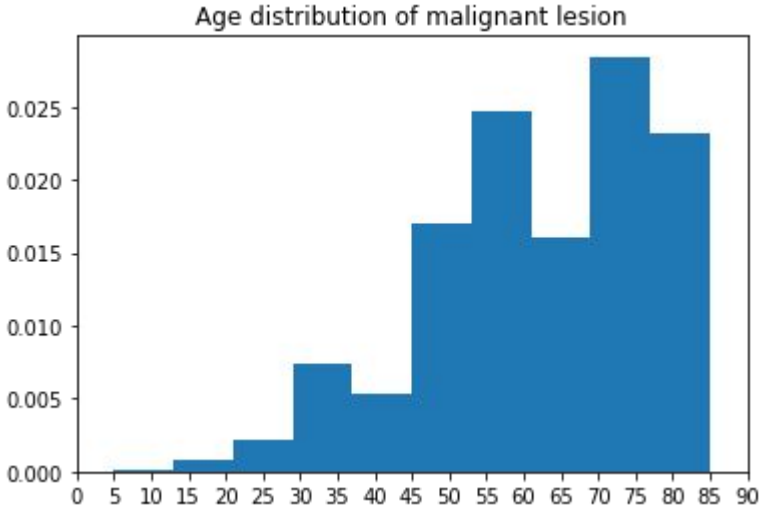



P value of independent T - test : 0

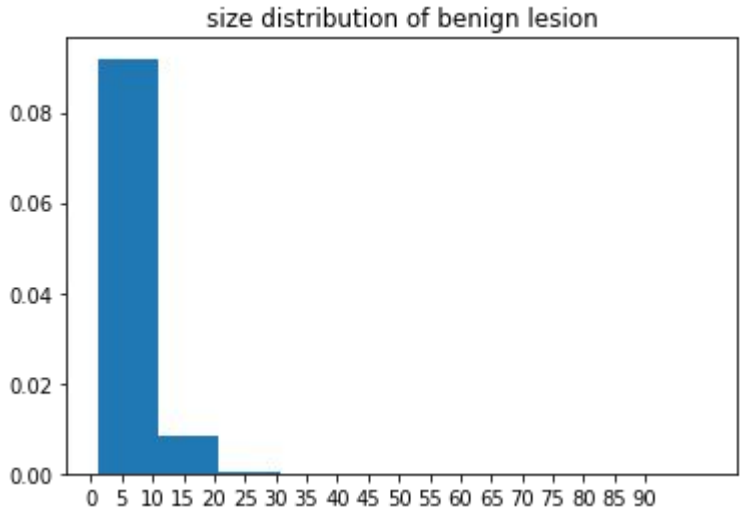
# Benign vs. Malignant: Age - w/o some data

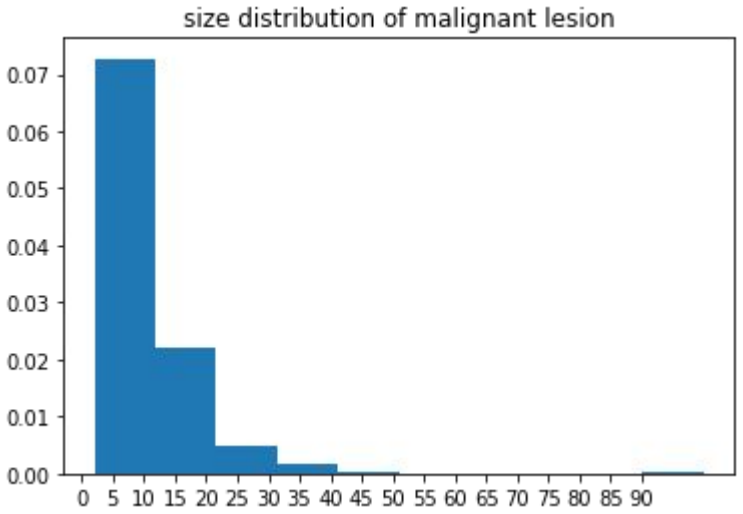



P value of independent T - test : 9.443773241442537e-276

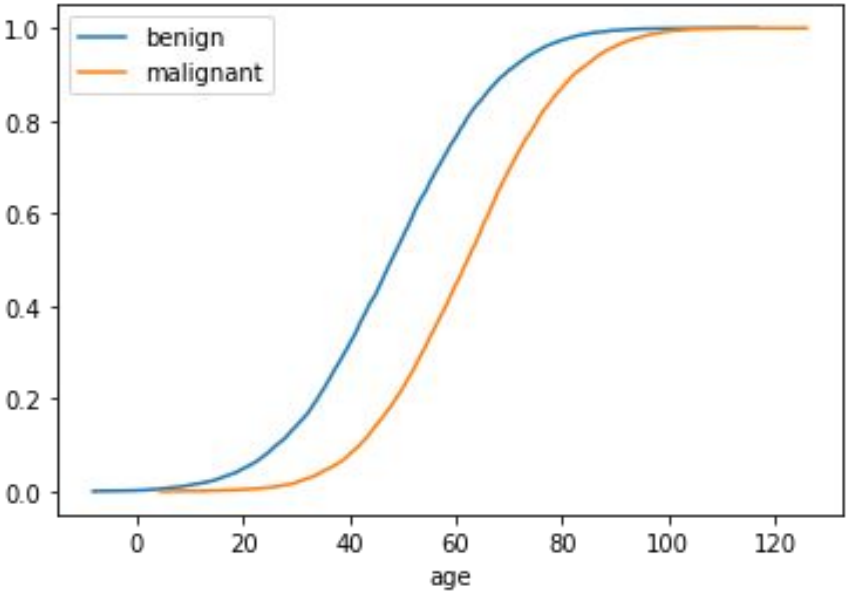
# Benign vs. Malignant: Size of the lesion

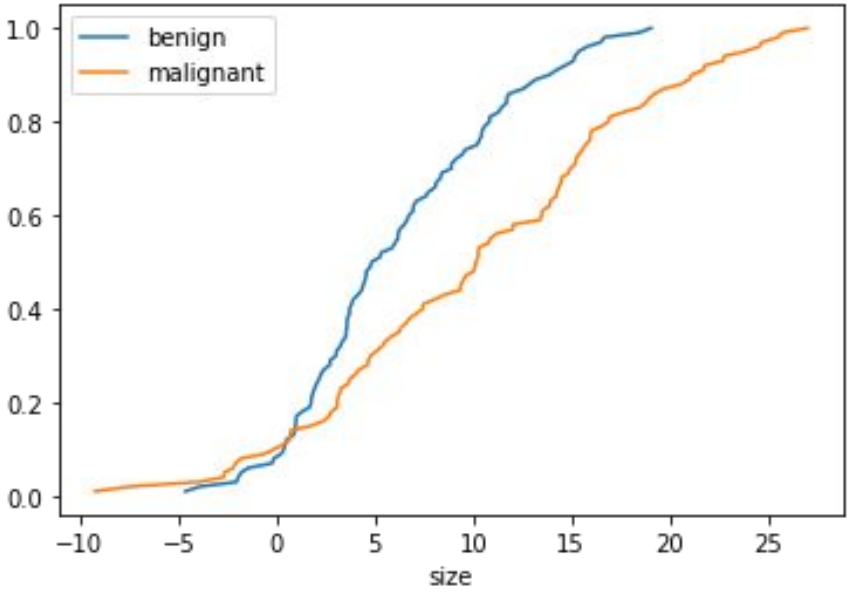


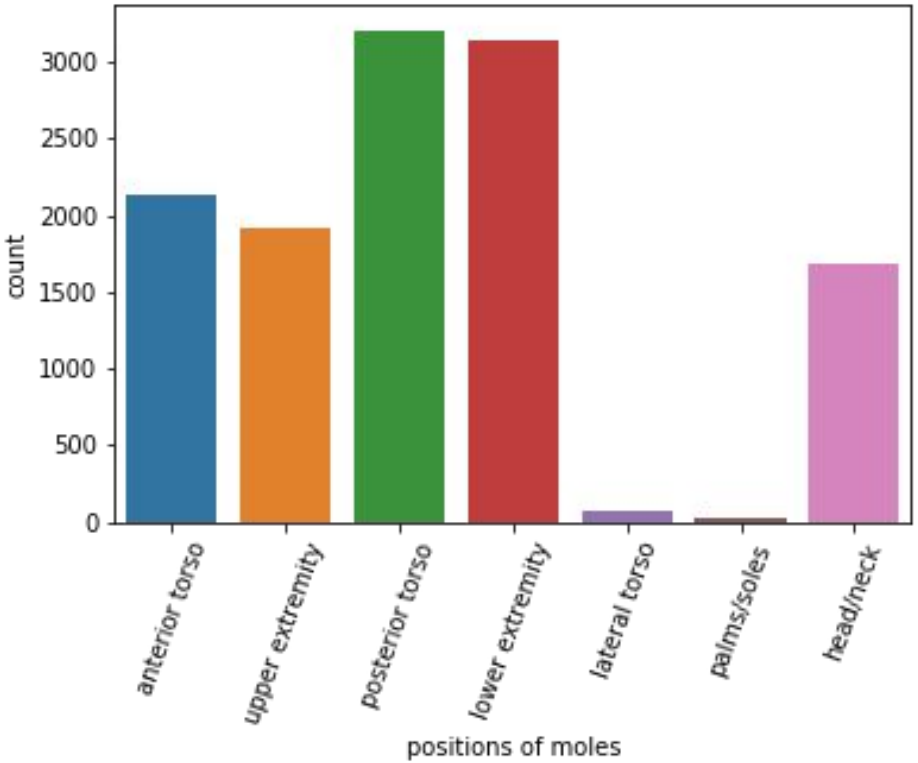

P value of independent T - test : 2.7279515204510347e-29

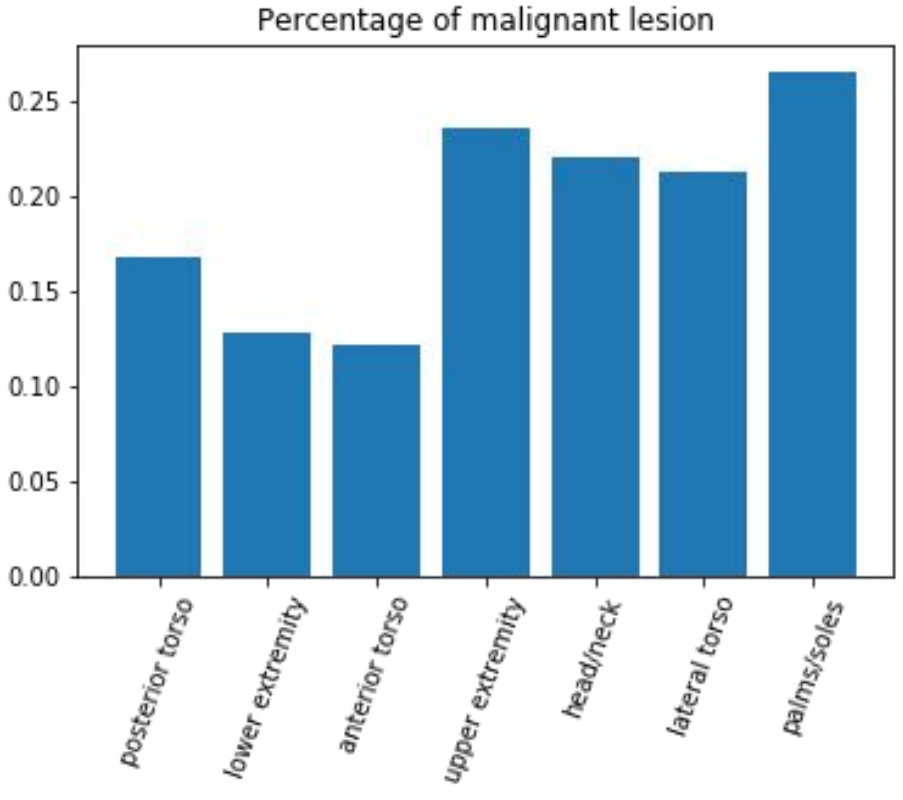
# ECDF of age range and size of lesion



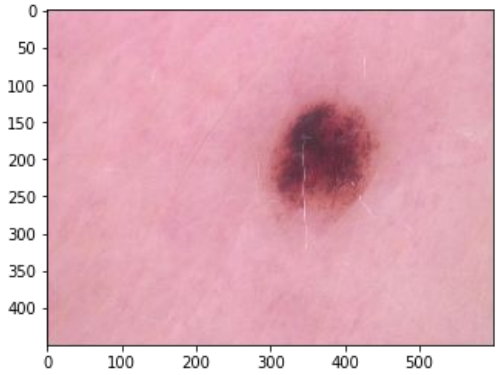



# Location of the lesion and percentage of malignancy



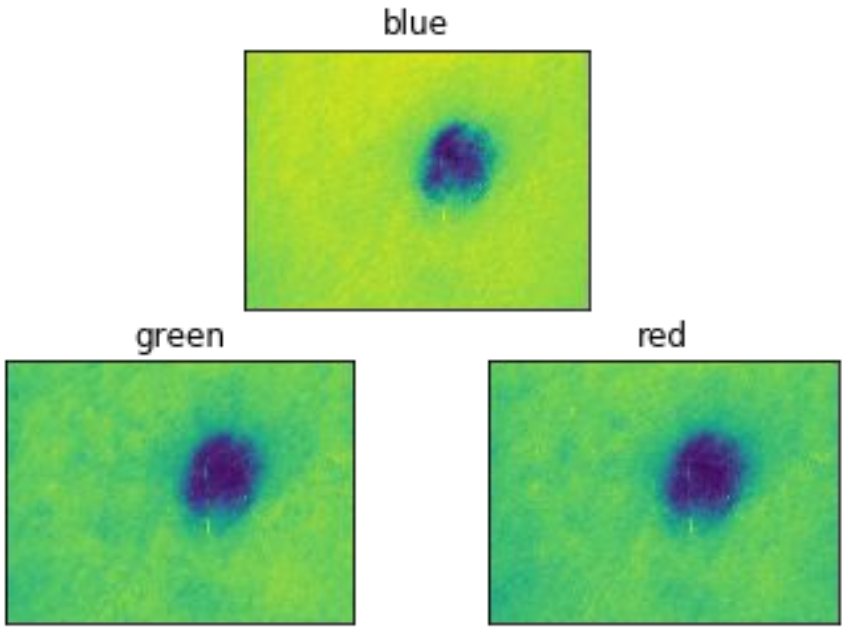



# Image preprocessing - normalize and color split



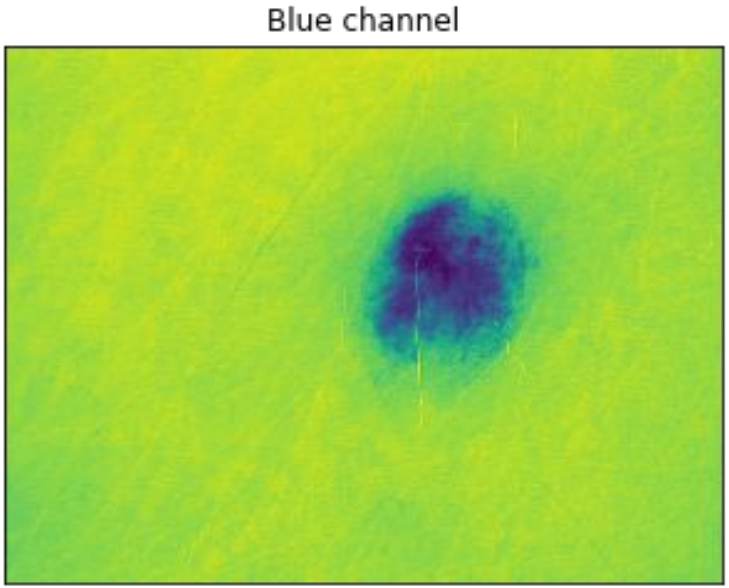

Original

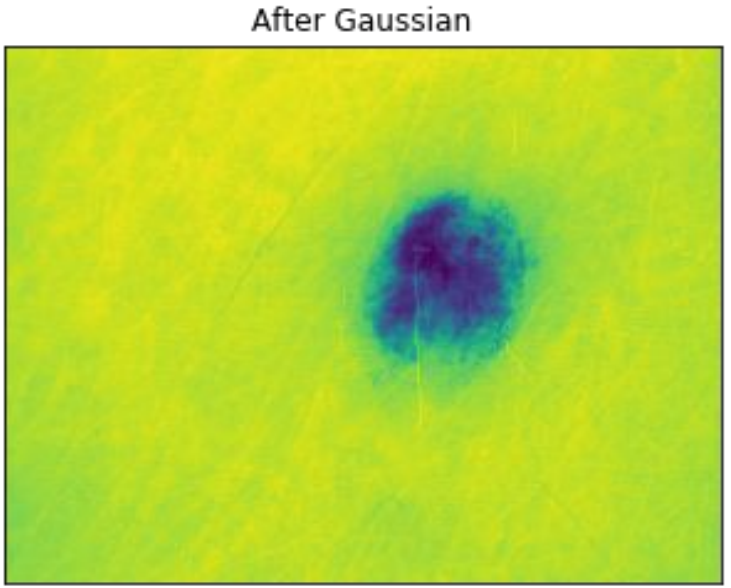

Normalized



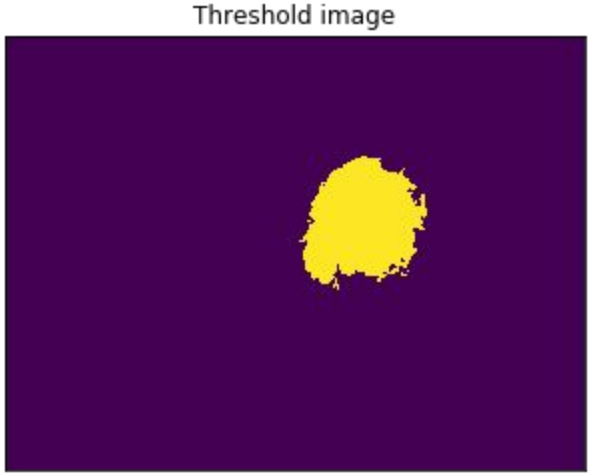

Color channel split

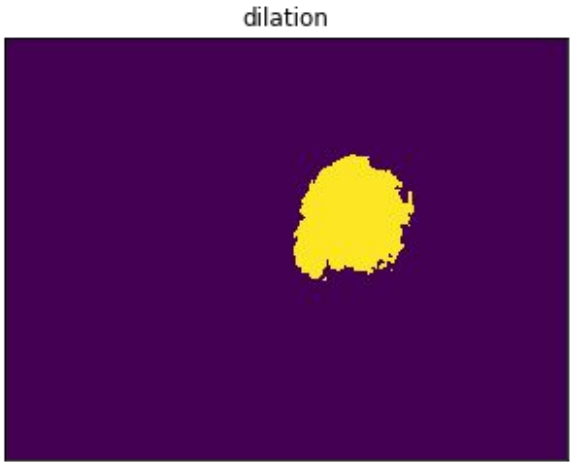
# Image preprocessing - Gaussian filter

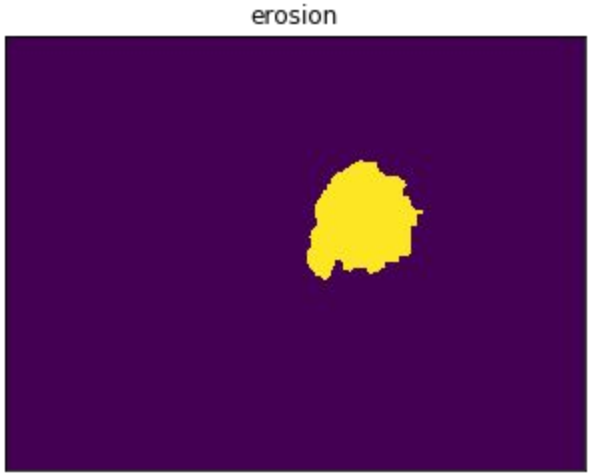


# Image preprocessing - Image thresholding




Binary + Otsu’s binarization

Kernel = 3 x 3

# Result of image segmentation

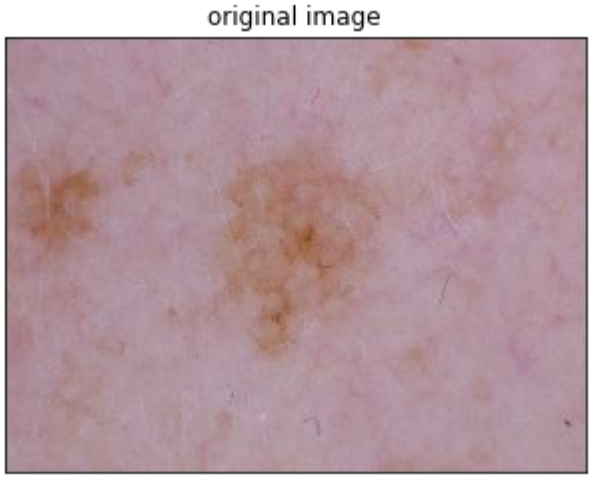
original image

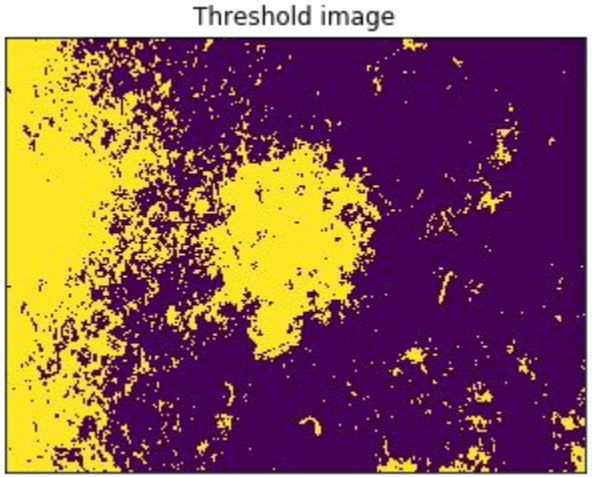
mask on image

# Problem with manual process of image segmentation

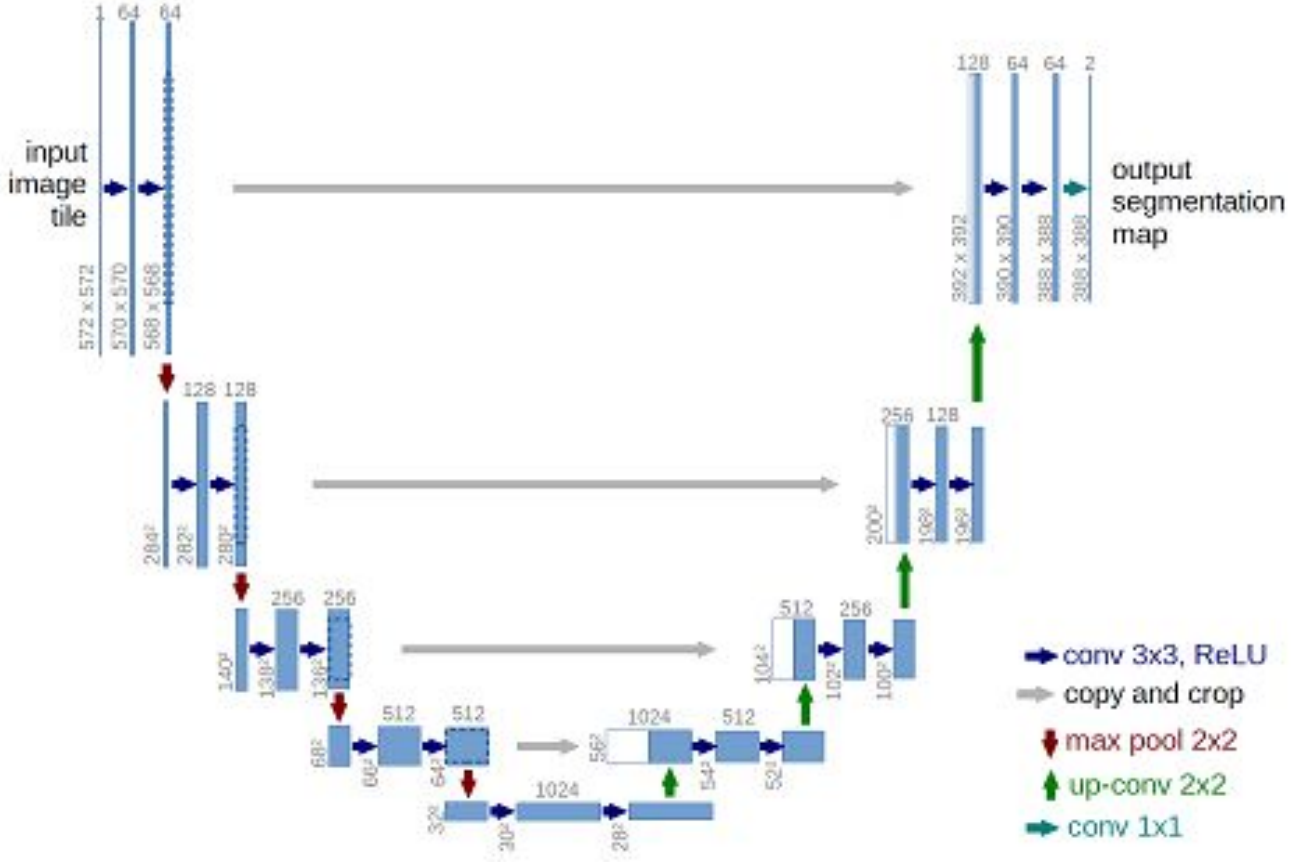
=> Very very slow

=> high chance of running into images that wouldn’t work

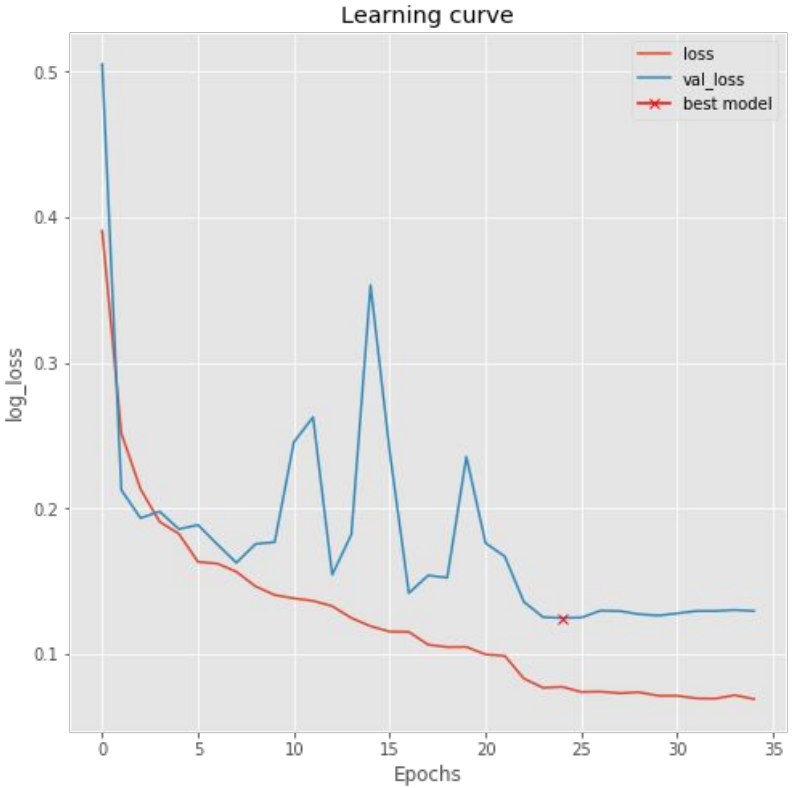


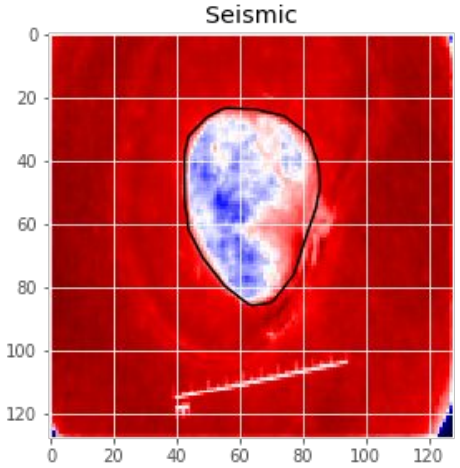



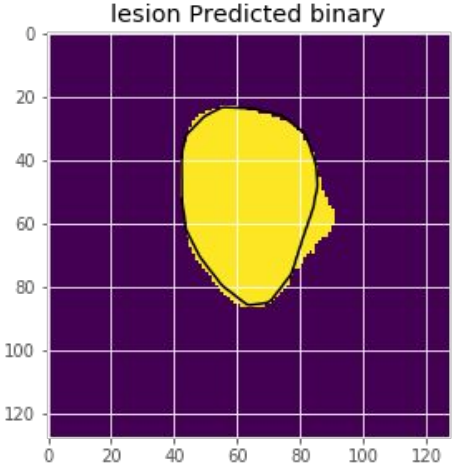

# Using UNET model for image segmentation



# Getting the best model



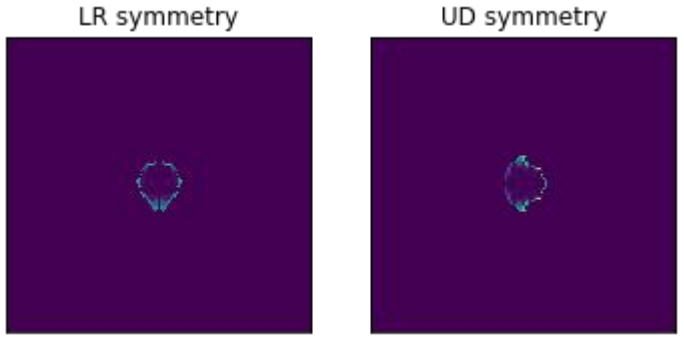



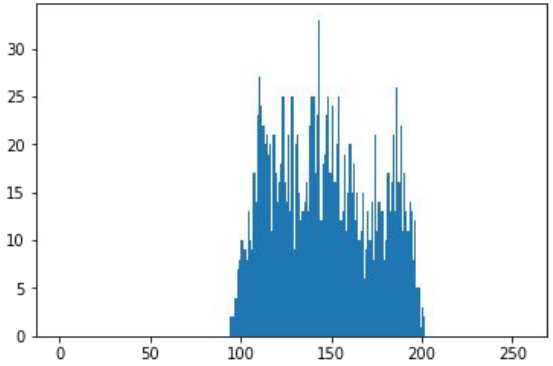
# ABC of ABCDE

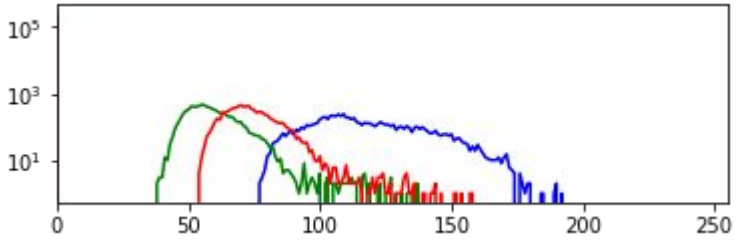
Asymmetry => tilt the lesion on major axis and compare the difference when flipped horizontally or vertically.

Border=> calculate gradient from cropped image of borders

Color => draw color historgram and get standard deviation

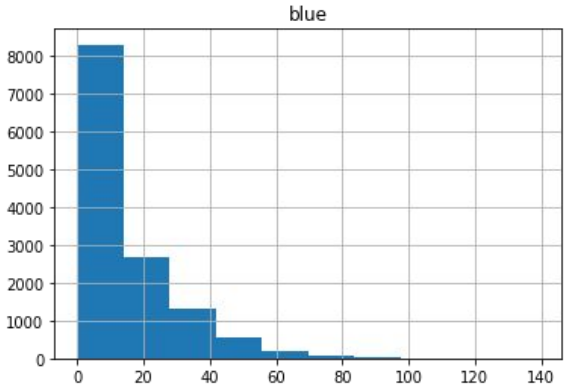


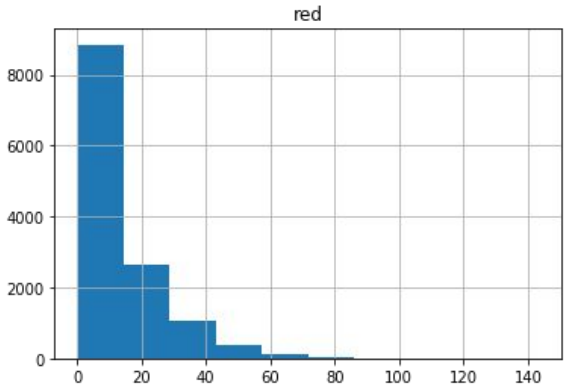



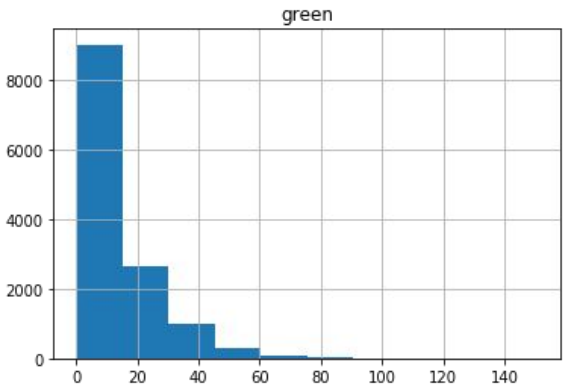



# Comparison between benign and malignant lesions

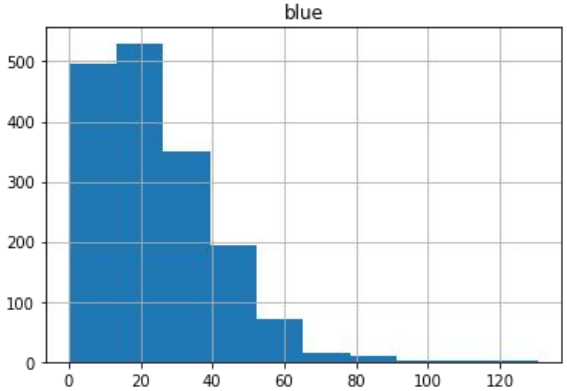
| | | symmetry | border | red | blue | green |
|---|---|---|---|---|---|---|
| malig | std | 3904182.5 | 0.560375 | 16.15623 | 16.63434 | 16.25676 |
| | mean | 815379.11 | 1.788012 | 21.0908 | 24.06696 | 21.85776 |
| beni | std | 1742113.5 | 0.544875 | 13.79489 | 14.79579 | 13.51443 |
| | mean | 241452.07 | 1.274017 | 12.59589 | 14.1669 | 12.67998 |

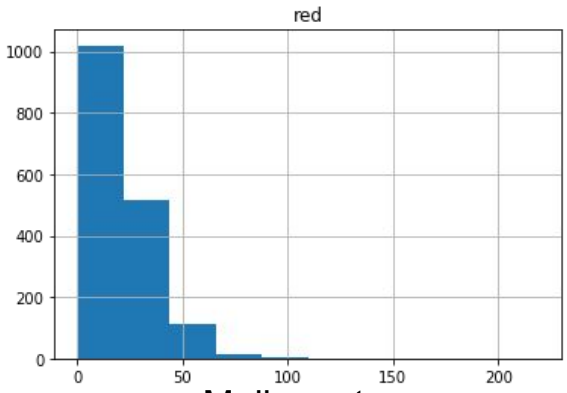


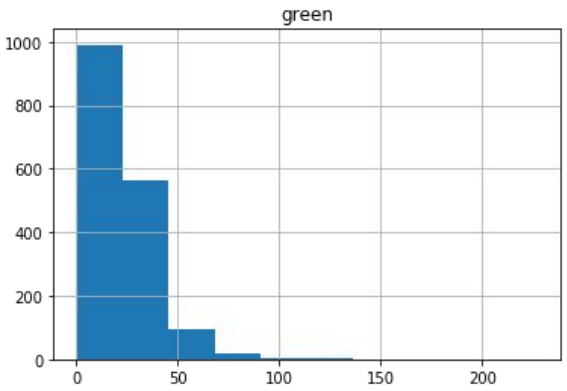



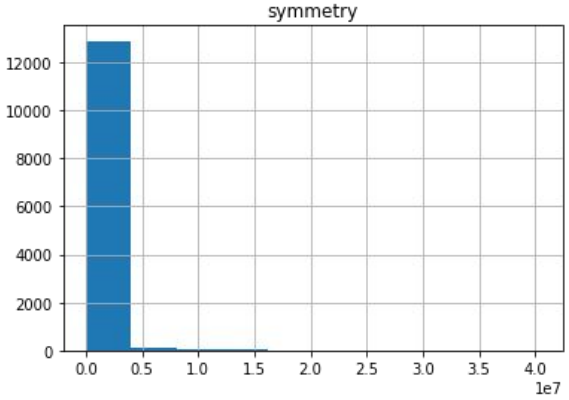

Benign

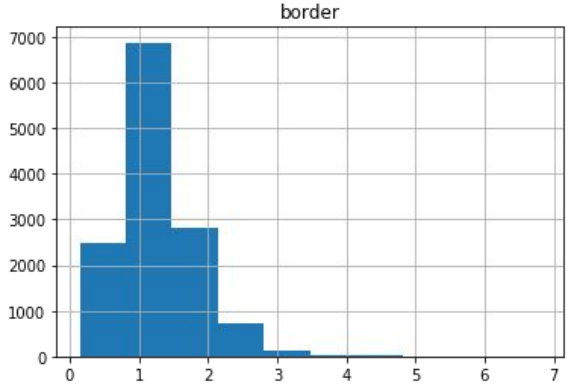


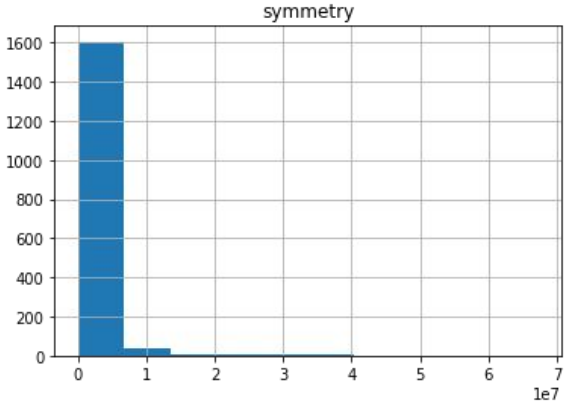



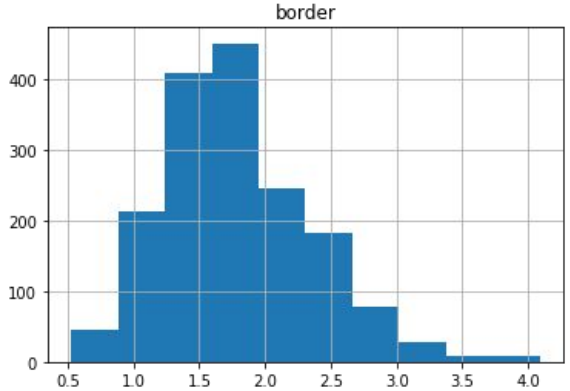

Malignant




Benign



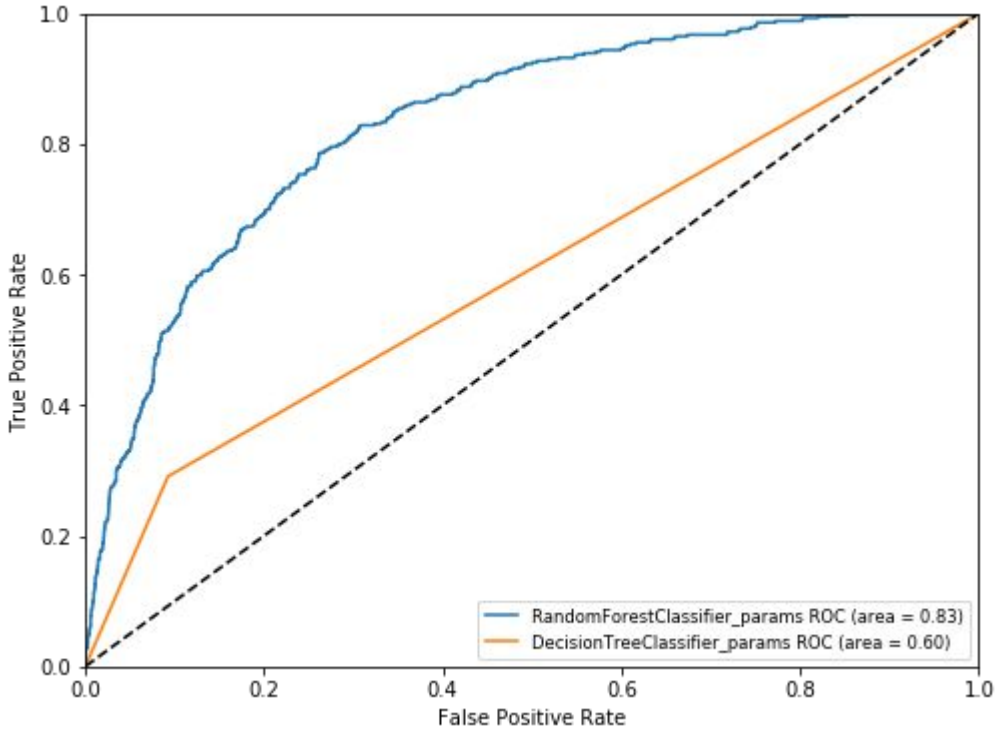

Malignant

# Features used for the model

- Symmetry
- Border
- Blue
- Red
- Green

=> A, B, C of ABCDE rule

# Random forest classification vs. Decision tree classification



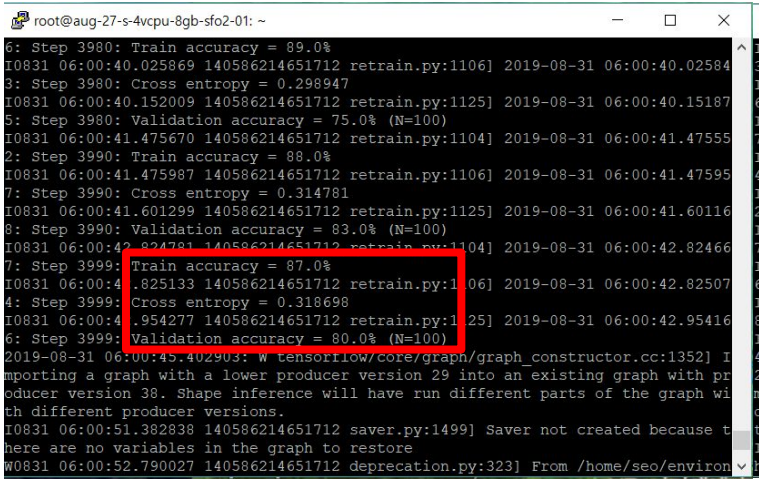

AUC of Random Forest Classifier 0.83
AUC of Decision Tree Classifier 0.60

# Transfer learning

- Retraining the last layer of a pre-trained CNN to classify the images between benign and malignant images
- Faster training
- Does not require much segmentation process

# Retraining steps

- Pretrained model: Inception V3 trained on ImageNet
- Training on top layer => training accuracy, validation accuracy, cross entropy
  - Cross entropy => loss function. Should be as small as possible




Resized => cross entropy 0.3187
Validation accuracy 80%

Regular+Resized => cross entropy 0.1160
Validation accuracy 90%

# Conclusion

- For the image processing problem, unless we know exactly how each image differs it is useful to use deep learning

More specific conclusion regarding things I tried and found