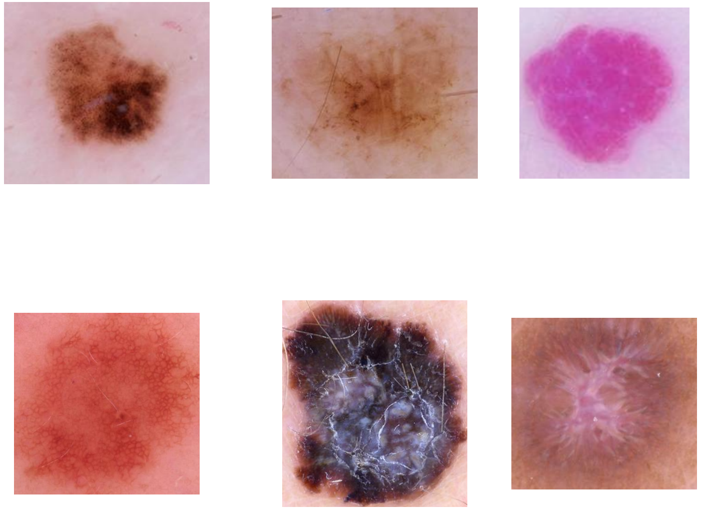
Capstone 3:
Image Classification

Jarom Hatch

Introduction

Problem Statement

- Machine learning is evolving and more and more becoming part of our everyday lives
  - Predicting stock prices
  - Giving tv show and product suggestions
  - Helping self driving cars identify people, vehicles, and obstacles
- Machine learning can and is also used in the medical world
- Can we use images, as well as some specific metadata associated with those images, to accurately diagnose skin lesions?
  - Beneficial for both patients and doctors to speed up diagnoses and treatment plans
  - Potentially easier and less expensive than biopsies

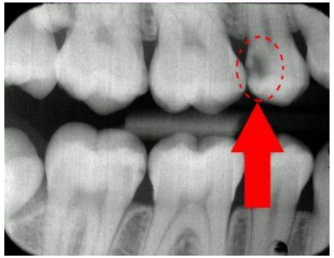


Target Audience

- Dermatologists and Patients with increased risk of skin disease
    - Although, the benefits from this learning can be applied to all aspects of medicine where visual identification of issues is present
- Secondary audience:
    - Physicians as a whole, Pharmaceuticals

Chickenpox rash

Smallpox rash



Data Overview

Initial Datasets

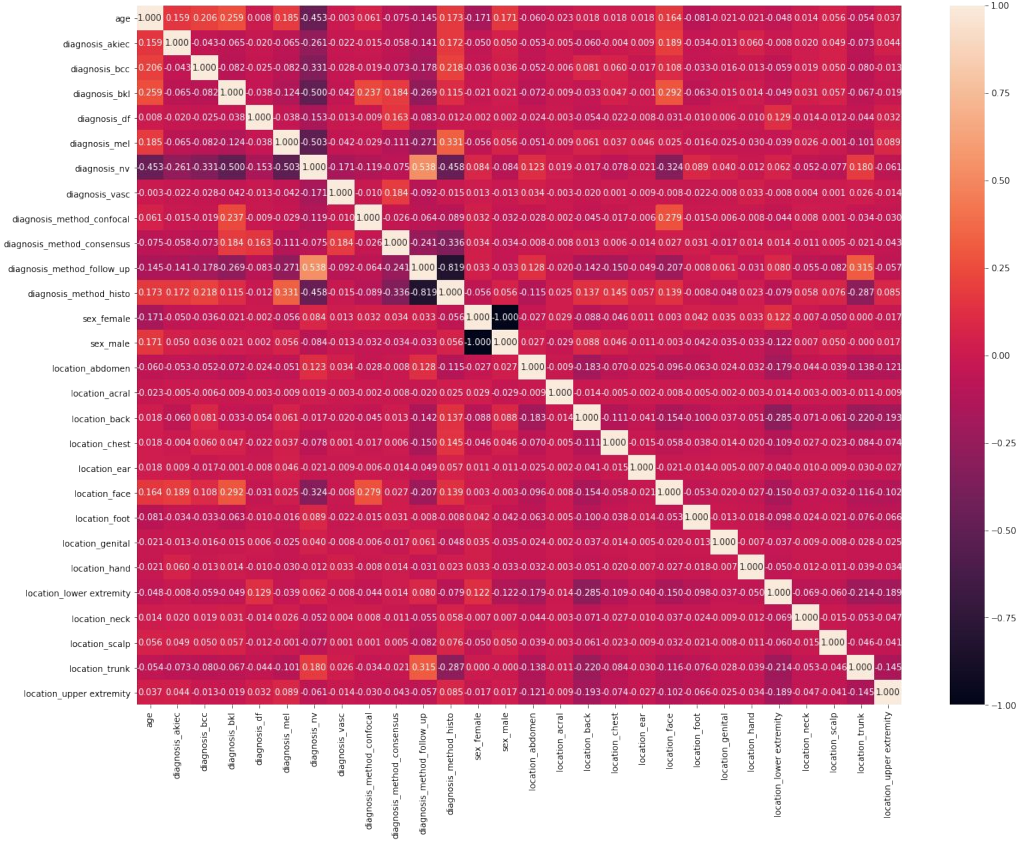
- **2** main types provided:
  - Images
  - Metadata
- **Images**:
  - 10,015 color images comprised of 7 different lesions
  - 450 x 600 pixels
- **Metadata**:
  - Lesion ID, Image ID, Diagnosis, Diagnosis method, age, sex, and location of lesions
    - 7 different diagnoses
    - 4 different diagnosis methods
    - Ages between 0 and 85
    - 14 locations
    - Male and Female patients

Data Wrangling

- The metadata was missing some values for different features that needed to be filled
  - All features were combined into one large feature to determine the most common values to fill missing data
- Images were initially all together in two files
  - Not optimal for Tensorflow Keras model
- Images were split based on diagnosis from the metadata and image ID into separate folders labelled by diagnosis

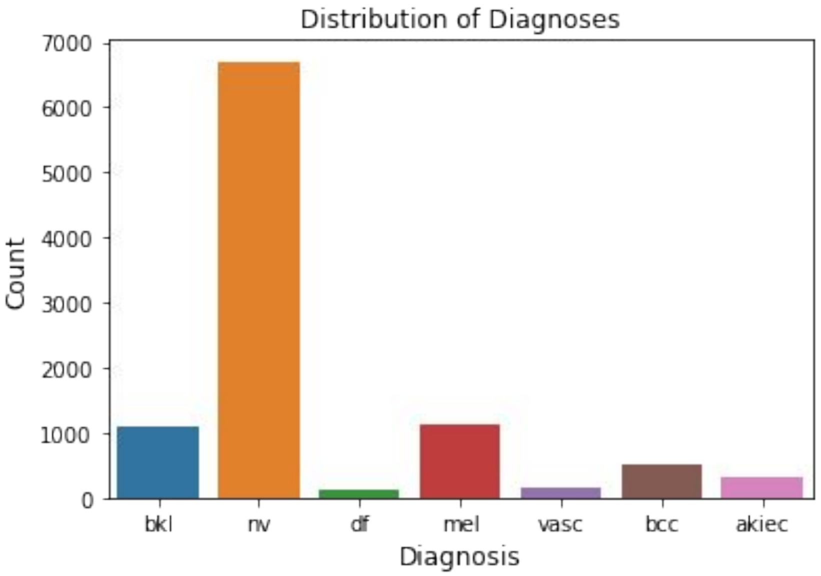
Exploratory Data Analysis

- The correlation matrix showed some interesting information
  - NV diagnosis had a high negative correlation with age and histopathology and a positive correlation with follow-up
  - The other diagnoses had a range of positive and negative correlations, although not as high as NV

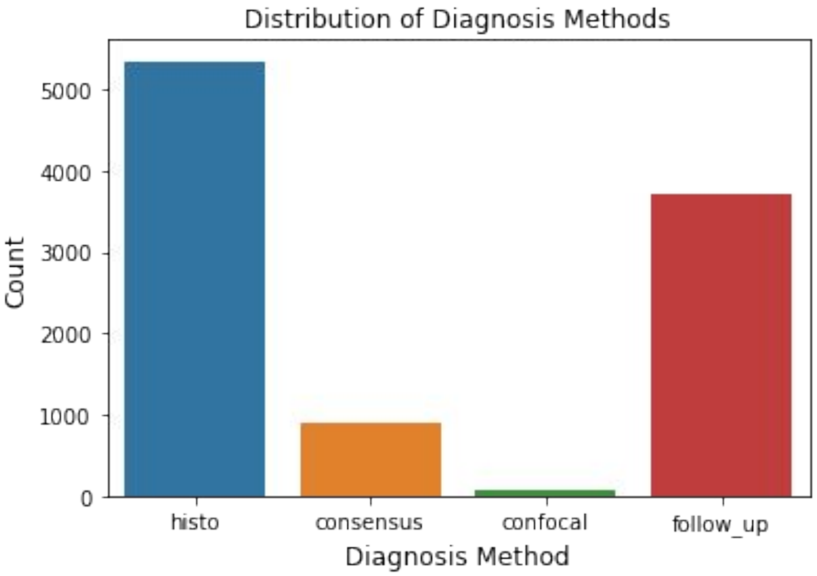


	age	diagnosis_akiec	diagnosis_bcc	diagnosis_bkl	diagnosis_df	diagnosis_mel	diagnosis_nv	diagnosis_vasc	diagnosis_method_confocal	diagnosis_method_consensus	diagnosis_method_follow_up	diagnosis_method_histo	sex_female	sex_male	location_abdomen	location_acral	location_back	location_chest	location_ear	location_face	location_foot	location_genital	location_hand	location_lower extremity	location_neck	location_scalp	location_trunk	location_upper extremity
age	1.000	0.159	0.206	0.259	0.008	0.185	-0.453	-0.003	0.061	-0.075	-0.145	0.173	-0.171	0.171	-0.060	-0.023	0.018	0.018	0.018	0.164	-0.081	-0.021	-0.021	-0.048	0.014	0.056	-0.054	0.037
diagnosis_akiec	0.159	1.000	-0.043	-0.065	-0.020	-0.065	-0.261	-0.022	-0.015	-0.058	-0.141	0.172	-0.050	0.050	-0.053	-0.005	-0.060	-0.004	0.009	0.189	-0.034	-0.013	0.060	-0.008	0.020	0.049	-0.073	0.044
diagnosis_bcc	0.206	-0.043	1.000	-0.082	-0.025	-0.082	-0.331	-0.028	-0.019	-0.073	-0.178	0.218	-0.036	0.036	-0.052	-0.006	0.081	0.060	-0.017	0.108	-0.033	-0.016	-0.013	-0.059	0.019	0.050	-0.080	-0.013
diagnosis_bkl	0.259	-0.065	-0.082	1.000	-0.038	-0.124	-0.500	-0.042	0.237	0.184	-0.269	0.115	-0.021	0.021	-0.072	-0.009	-0.033	0.047	-0.001	0.292	-0.063	-0.015	0.014	-0.049	0.031	0.057	-0.067	-0.019
diagnosis_df	0.008	-0.020	-0.025	-0.038	1.000	-0.038	-0.153	-0.013	-0.009	0.163	-0.083	-0.012	-0.002	0.002	-0.024	-0.003	-0.054	-0.022	-0.008	-0.031	-0.010	0.006	-0.010	0.129	-0.014	-0.012	-0.044	0.032
diagnosis_mel	0.185	-0.065	-0.082	-0.124	-0.038	1.000	-0.503	-0.042	-0.029	-0.111	-0.271	0.331	-0.056	0.056	-0.051	-0.009	0.061	0.037	0.046	0.025	-0.016	-0.025	-0.030	-0.039	0.026	-0.001	-0.101	0.089
diagnosis_nv	-0.453	-0.261	-0.331	-0.500	-0.153	-0.503	1.000	-0.171	-0.119	-0.075	0.538	-0.458	0.084	-0.084	0.123	0.019	-0.017	-0.078	-0.021	-0.324	0.089	0.040	-0.012	0.062	-0.052	-0.077	0.180	-0.061
diagnosis_vasc	-0.003	-0.022	-0.028	-0.042	-0.013	-0.042	-0.171	1.000	-0.010	0.184	-0.092	-0.015	0.013	-0.013	0.034	-0.003	-0.020	0.001	-0.009	-0.008	-0.022	-0.008	0.033	-0.008	0.004	0.001	0.026	-0.014
diagnosis_method_confocal	0.061	-0.015	-0.019	0.237	-0.009	-0.029	-0.119	-0.010	1.000	-0.026	-0.064	-0.089	0.032	-0.032	-0.028	-0.002	-0.045	-0.017	-0.006	0.279	-0.015	-0.006	-0.008	-0.044	0.008	0.001	-0.034	-0.030
diagnosis_method_consensus	-0.075	-0.058	-0.073	0.184	0.163	-0.111	-0.075	0.184	-0.026	1.000	-0.241	-0.336	0.034	-0.034	-0.008	-0.008	0.013	0.006	-0.014	0.027	0.031	-0.017	0.014	0.014	-0.011	0.005	-0.021	-0.043
diagnosis_method_follow_up	-0.145	-0.141	-0.178	-0.269	-0.083	-0.271	0.538	-0.092	-0.064	-0.241	1.000	-0.819	0.033	-0.033	0.128	-0.020	-0.142	-0.150	-0.049	-0.207	-0.008	0.061	-0.031	0.080	-0.055	-0.082	0.315	-0.057
diagnosis_method_histo	0.173	0.172	0.218	0.115	-0.012	0.331	-0.458	-0.015	-0.089	-0.336	-0.819	1.000	-0.056	0.056	-0.115	0.025	0.137	0.145	0.057	0.139	-0.008	-0.048	0.023	-0.079	0.058	0.076	-0.287	0.085
sex_female	-0.171	-0.050	-0.036	-0.021	-0.002	-0.056	0.084	0.013	0.032	0.034	0.033	-0.056	1.000	-1.000	-0.027	0.029	-0.088	-0.046	0.011	0.003	0.042	0.035	0.033	0.122	-0.007	-0.050	0.000	-0.017
sex_male	0.171	0.050	0.036	0.021	0.002	0.056	-0.084	-0.013	-0.032	-0.034	-0.033	0.056	-1.000	1.000	0.027	-0.029	0.088	0.046	-0.011	-0.003	-0.042	-0.035	-0.033	-0.122	0.007	0.050	-0.000	0.017
location_abdomen	-0.060	-0.053	-0.052	-0.072	-0.024	-0.051	0.123	0.034	-0.028	-0.008	0.128	-0.115	-0.027	0.027	1.000	-0.009	-0.183	-0.070	-0.025	-0.096	-0.063	-0.024	-0.032	-0.179	-0.044	-0.039	-0.138	-0.121
location_acral	-0.023	-0.005	-0.006	-0.009	-0.003	-0.009	0.019	-0.003	-0.002	-0.008	-0.020	0.025	0.029	-0.029	-0.009	1.000	-0.014	-0.005	-0.002	-0.008	-0.005	-0.002	-0.003	-0.014	-0.003	-0.003	-0.011	-0.009
location_back	0.018	-0.060	0.081	-0.033	-0.054	0.061	-0.017	-0.020	-0.045	0.013	-0.142	0.137	-0.088	0.088	-0.183	-0.014	1.000	-0.111	-0.041	-0.154	-0.100	-0.037	-0.051	-0.285	-0.071	-0.061	-0.220	-0.193
location_chest	0.018	-0.004	0.060	0.047	-0.022	0.037	-0.078	0.001	-0.017	0.006	-0.150	0.145	-0.046	0.046	-0.070	-0.005	-0.111	1.000	-0.015	-0.058	-0.038	-0.014	-0.020	-0.109	-0.027	-0.023	-0.084	-0.074
location_ear	0.018	0.009	-0.017	-0.001	-0.008	0.046	-0.021	-0.009	-0.006	-0.014	-0.049	0.057	0.011	-0.011	-0.025	-0.002	-0.041	-0.015	1.000	-0.021	-0.014	-0.005	-0.007	-0.040	-0.010	-0.009	-0.030	-0.027
location_face	0.164	0.189	0.108	0.292	-0.031	0.025	-0.324	-0.008	0.279	0.027	-0.207	0.139	0.003	-0.003	-0.096	-0.008	-0.154	-0.058	-0.021	1.000	-0.053	-0.020	-0.027	-0.150	-0.037	-0.032	-0.116	-0.102
location_foot	-0.081	-0.034	-0.033	-0.063	-0.010	-0.016	0.089	-0.022	-0.015	0.031	-0.008	-0.008	0.042	-0.042	-0.063	-0.005	-0.100	-0.038	-0.014	-0.053	1.000	-0.013	-0.018	-0.098	-0.024	-0.021	-0.076	-0.066
location_genital	-0.021	-0.013	-0.016	-0.015	0.006	-0.025	0.040	-0.008	-0.006	-0.017	0.061	-0.048	0.035	-0.035	-0.024	-0.002	-0.037	-0.014	-0.005	-0.020	-0.013	1.000	-0.007	-0.037	-0.009	-0.008	-0.028	-0.025
location_hand	-0.021	0.060	-0.013	0.014	-0.010	-0.030	-0.012	0.033	-0.008	0.014	-0.031	0.023	0.033	-0.033	-0.032	-0.003	-0.051	-0.020	-0.007	-0.027	-0.018	-0.007	1.000	-0.050	-0.012	-0.011	-0.039	-0.034
location_lower extremity	-0.048	-0.008	-0.059	-0.049	0.129	-0.039	0.062	-0.008	-0.044	0.014	0.080	-0.079	0.122	-0.122	-0.179	-0.014	-0.285	-0.109	-0.040	-0.150	-0.098	-0.037	-0.050	1.000	-0.069	-0.060	-0.214	-0.189
location_neck	0.014	0.020	0.019	0.031	-0.014	0.026	-0.052	0.004	0.008	-0.011	-0.055	0.058	-0.007	0.007	-0.044	-0.003	-0.071	-0.027	-0.010	-0.037	-0.024	-0.009	-0.012	-0.069	1.000	-0.015	-0.053	-0.047
location_scalp	0.056	0.049	0.050	0.057	-0.012	-0.001	-0.077	0.001	0.001	0.005	-0.082	0.076	-0.050	0.050	-0.039	-0.003	-0.061	-0.023	-0.009	-0.032	-0.021	-0.008	-0.011	-0.060	-0.015	1.000	-0.046	-0.041
location_trunk	-0.054	-0.073	-0.080	-0.067	-0.044	-0.101	0.180	0.026	-0.034	-0.021	0.315	-0.287	0.000	-0.000	-0.138	-0.011	-0.220	-0.084	-0.030	-0.116	-0.076	-0.028	-0.039	-0.214	-0.053	-0.046	1.000	-0.145
location_upper extremity	0.037	0.044	-0.013	-0.019	0.032	0.089	-0.061	-0.014	-0.030	-0.043	-0.057	0.085	-0.017	0.017	-0.121	-0.009	-0.193	-0.074	-0.027	-0.102	-0.066	-0.025	-0.034	-0.189	-0.047	-0.041	-0.145	1.000

Exploratory Data Analysis (cont.)

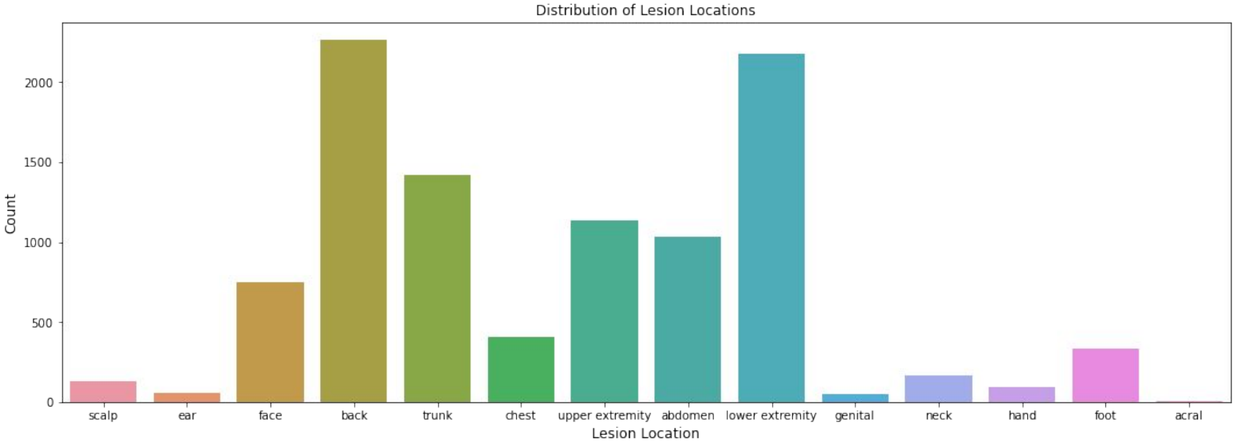


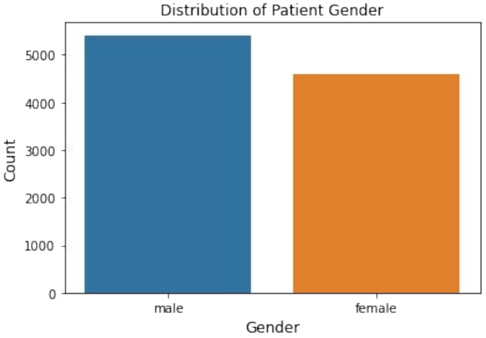
Distribution of Diagnoses



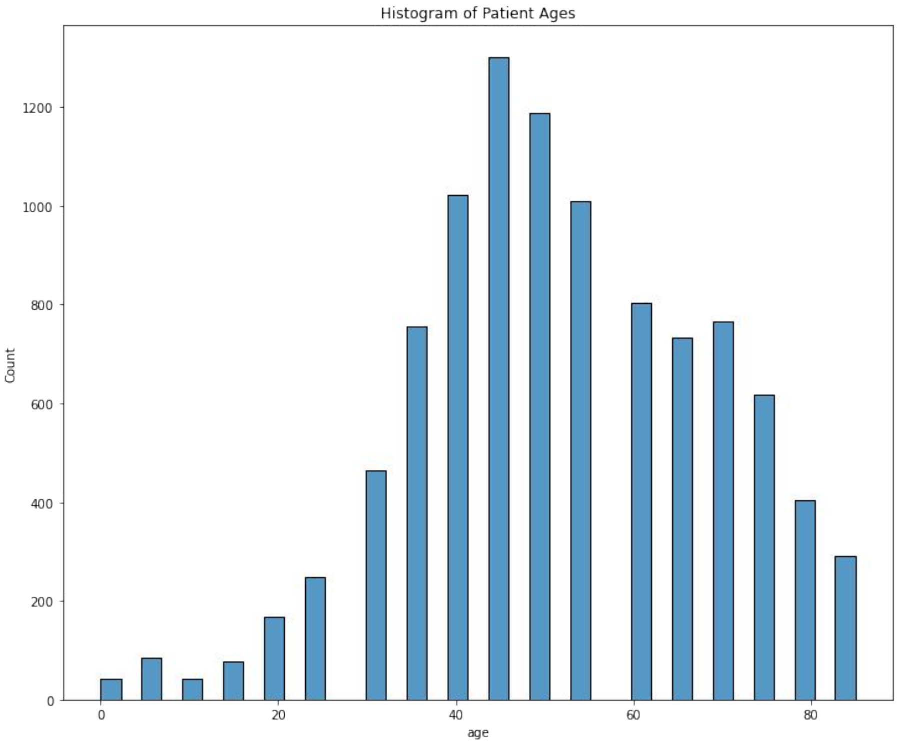
Distribution of Diagnosis Methods

Exploratory Data Analysis (cont.)





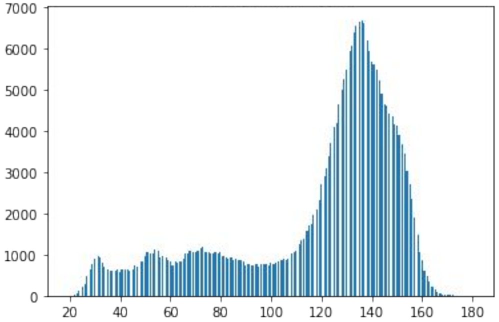
Exploratory Data Analysis (cont.)

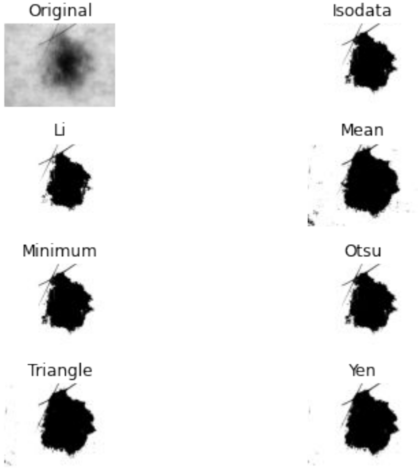


Basic Image Analysis

- Another goal of this project was to determine if the images were ready for machine learning and if any interesting information could be gleaned from them as part of the EDA process
- Different processing techniques available through Python libraries were used for this section
- This included:
  - Slicing images by color channels
  - Thresholding
  - Edge detection
  - Exposure adjustment
  - Contour detection

Basic Image Analysis


Blue Channel Levels


Original

Isodata

Li

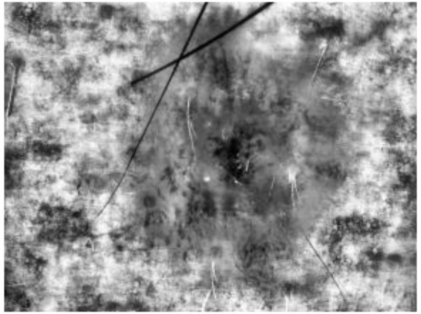
Mean

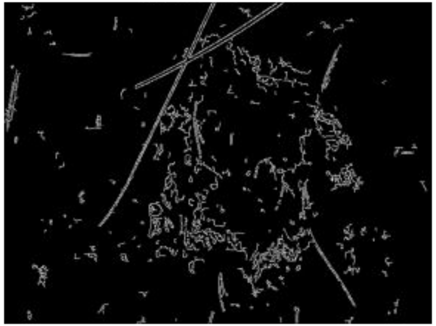
Minimum

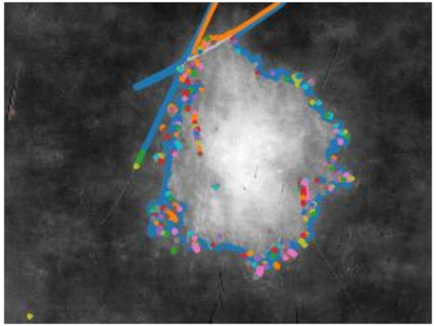
Otsu

Triangle

Yen


Adaptive Histogram Equalized


Edges with Canny


Contours

Basic Image Analysis

- Brought out edge detail as well as underlying surface detail on the skin of the patient.
    - Useful in differentiating between different lesion types
    - Visualizing skin damage not initially visible
- Different lesion types also have different colors and similar shapes
    - Probably not useful to edit the images from their original format

Machine Learning

Machine Learning Setup

- Two types of data were used
    - Images
    - Tabular data
- Images were scaled to prepare for machine learning
- Tabular data was optimized using encoding to ensure numerical values
    - Data not relevant to the analysis was dropped
- Machine learning models chosen from typically used models for multi-class classification
    - Images
        - Sequential Keras model with multiple layers
            - Rescaling, 2D Convolutional layers, 2D MaxPooling layers, Dropout layer to prevent overfitting, Flatten layer, and Dense layer
    - Tabular Data
        - Random Forest Classifier
        - Easy Ensemble Classifier
        - KNN Classifier
        - Naive Bayes Classifier
- Image model assessed on accuracy and loss
- Tabular data model assessed on accuracy and classification report metrics

Machine Learning Metrics

Metrics from best image model:

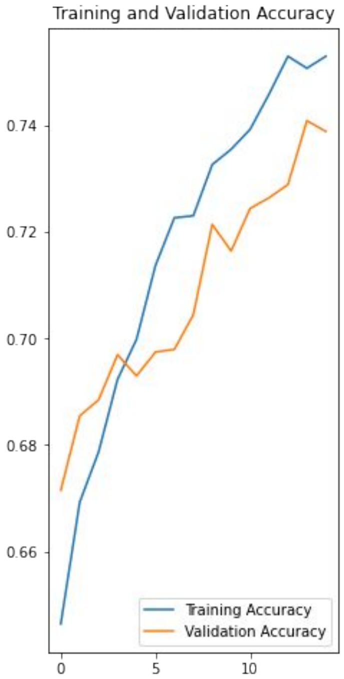
Model	Accuracy	Loss
Keras Sequential	73%	0.75

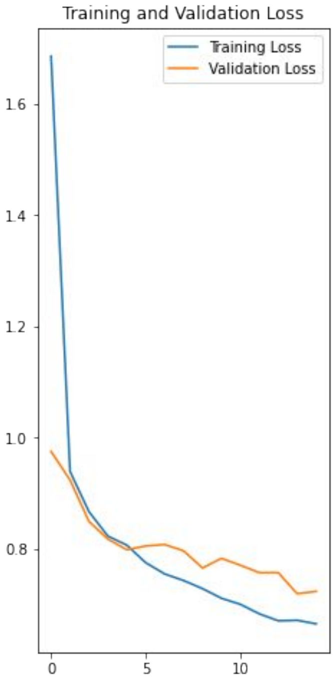
Metrics from best tabular data model:

Model	Accuracy
Random Forest Classifier	73%

Keras Sequential Model

- The best image classification model utilized the following features:
    - Data augmentation
        - Random flips
        - Random rotation
        - Random zoom
    - 11 additional layers, including:
        - Rescaling
        - Convolutional 2D
        - MaxPooling 2D
        - Dropout
        - Flatten
        - Dense
    - Adam optimizer
    - Sparse categorical cross entropy loss
    - Accuracy metric
    - 15 Epochs
    - 128 Batch Size





Random Forest Classifier

- Best tabular data model utilized a generic random forest classifier
    - 200 Estimators
    - Accuracy metric
        - 73%
    - Classification report

Random Forest Classifier

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

Conclusion

Conclusions

- Machine learning was able to work for both classification using image and text data
- Both models produced very similar accuracy results
- Since the text data had distinct features, we were able to pull out metrics for each classification
    - Based on the precision, recall, and F1 scores, the classes were very difficult to predict from the provided datasets
    - Even after running cross validation, the accuracy didn’t increase
    - Our best metrics came from classes which had the highest instances, meaning more data to analyze
- Moving forward, more precise fine tuning may be necessary to adjust for the class imbalances in our modeling
- Experiment with combining the methods in a wide and deep model