Name NetID

Yuan Shi yuanshi4

Meng Meng mmeng3

Siyuan Huang shuang99

Skin Cancer Diagnostics Report

Section 1: Project description and summary

Image analysis is becoming one of the leading trends in statistical learning. Especially, there are a lot of well-constructed techniques developed for image analysis such as openCV, Pillow and SciPy in Python. As for practical applications, image analysis is extremely helpful in assisting doctors' diagnosis of skin cancer: for example, accurately distinguishing benign and malignant moles will help identifying skin cancer at an early stage and preventing worsen cases. While now the diagnosing process mainly relies on doctors' domain knowledge and prior experience, we want to implement machine learning techniques and create several classification models for predicting the presence and absence of cancerous malignant cells, thus to provide more meaningful insights and help doctors getting more accurate diagnosis.

In the first part of our report, we will load the images into pixels, and convert them to RGB values as predictors for modeling. With pixels as input, Logistic Regression, Random Forest, and Support Vector Machine (SVM) classifiers will be utilized. Accuracy rates for the three classifiers are 66%, 72.2% and 73.3%, respectively. After parameter tuning and dimension reduction by PCA, accuracy rates of Logistic Regression was increased to 70% and 74.9% for Random Forest classifier.

In the next part, after reviewing the literature on both clinical and statistical fields, feature engineering was conducted and new features including areas and perimeters of moles, mean and standard deviation of color groups are added as predictors. Classification models in previous part are deployed using new features. Compared to previous modeling results, the accuracy rates of Random Forest and SVM models were improved to 77.8% and 82.2%, respectively.

Last but not least, we will compare our models based on different feature selections, and discuss possible future enhancements in terms of model performance.

Section 2: Data Pre-Processing

Before constructing classification models, we have done several steps to pre-process the image data:

1. Obtaining file paths for both benign and malignant image groups, each of them have 150 images and were stored in the working space. Some images with black frame were manually edited before converting to pixels.
2. Most of the images had an approximately 3:4 image aspect ratio and sizes of 1050 x 2500. In terms of pixels extraction, if we process the original image input, for every color image we would have width*height*3 "data points" for all pixels – this number is greater

than 7,000,000, which means we will have over 7 million features encoded as data columns.
    a. In order to the huge dimension, all images were scaled to 300 x 400 matching the 3:4 aspect ratio.
    b. After image resizing, we extracted pixels and flatten them with three color layers (R/G/B) to a vector which has dimensions of 1 x (300*400*3)
    c. This step brought our dataset to have 300 observations and 360,000 variables.
3. After fitting preliminary models using all 360,000 features, we noticed that pixels representing skin portion (with lighter color) would bring some noise to modeling.
    a. Features representing mole portion were manually selected from column 117,000 to 225,000.
    b. Current dataset had 300 observations and 108,000 variables.
4. Furthermore, before first and second cycle of modeling, we utilized PCA on current features and reduced total number of features to 20 principal components.
5. Training and testing sets were split using splitting ratio of 0.30 – this step was conducted three times for each cycle of modeling.

Section 3: Classification Models with Pixels

Three classification models were selected: Logistic Regression, Random Forest (RF), and Support Vector Machine (SVM) with radial basis kernel. In terms of hyperparameter tuning and improving model performance, we investigated the following variations:

- L1 and L2 regularization for Logistic Regression;
- Grid Search CV for max_depth (i.e. number of trees) and number of features for Random Forest;
- Different types of kernels and Grid Search CV for other parameters on Kernel SVM.

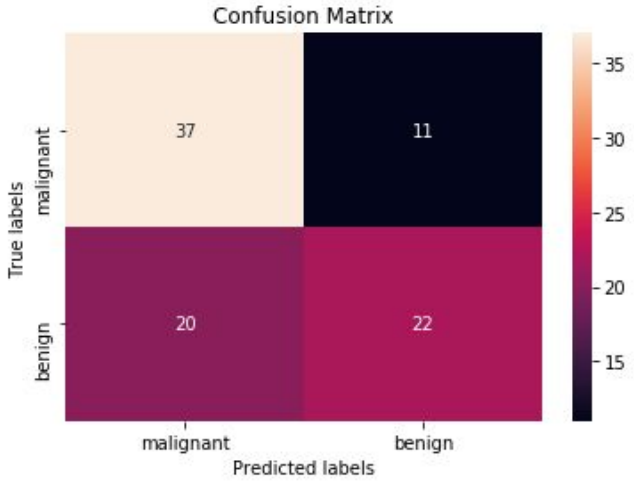
Logistic Regression

Logistics regression is a widely used model which examines the relationship between a binary outcome and predictor variables. For example, the presence or absence of skin cancer may be predicted by the RGB value of certain pixels from a patient's picture. The advantages of using logistic regression include being more efficient because it does not require many computational resources and being highly interpretable. Here in our model, the response variable uses 0 and 1 to denote "benign" and "malignant".

After we fit the predictors in the logistic regression model, we got the accuracy of 64% on the testing set. In order to improve the performance of this model, we standardized the dataset and fit the model again, this time we increased the accuracy to 66%.

Accuracy before standardized	0.64
Accuracy after standardized	0.66

The confusion matrix indicates the number of predicted labels vs true labels.



One more step forward, we applied PCA on normalized features and fit logistic regression classifier again with the 20 principal components that captured 90% of the total variation.

Prior to parameter tuning, the accuracy of this cycle goes up to 70%. Compared to the previous results, the FN value decreases to 16 from 20 out of 90 testing observations.

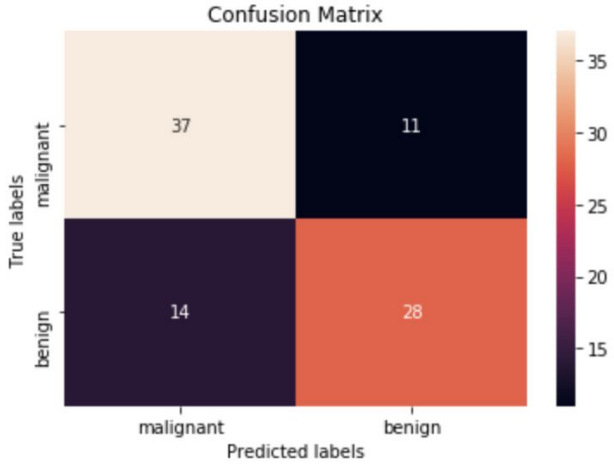
After tuning model with Grid Search CV using 10-fold cross validation, L1 regularization was selected and the final accuracy for logistic classifier became 0.69.

Random Forest Classifier

The Random Forest model is based on the idea of decision tree model, we sampling many decision trees with different node (p)m, and such tree models are able to provide us reliable prediction performance. Random forest can handle both regression and classification problems, in this project, we are trying to optimize the random forest classification model.

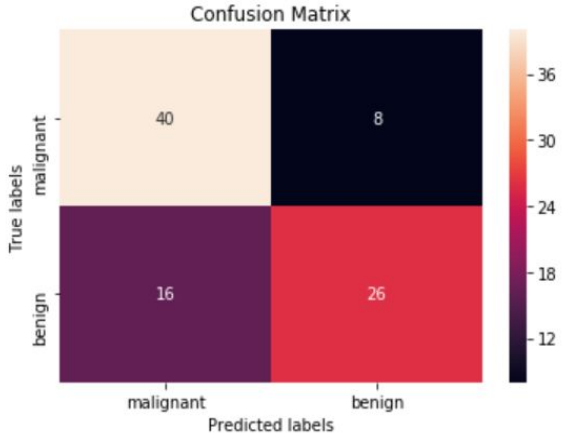
Some benefits of using random forest: Random forest is able to provide a reliable variable importance features, and it can take consideration of out-of-bag errors in order to prevent overfitting.

In this problem, we used the GridSearch feature in Python to find the optimal tuning parameter of random forest. Using the pixel as the predictor, and the parameters selected from the gridSearch, the random forest model provides an accuracy of 72.22% on the Test Set (30% of images) as shown in the confusion matrix below.



Kernel SVM Classifier

SVM is a supervised learning algorithm that plays the kernel trick to transform data input and find an optimal boundary between the possible outcomes based on these transformation. In the first cycle of model fitting, both linear kernel and non-linear kernel was used on the complete features set (360,000 columns). Given the default cost (C = 1) and gamma ('auto' option), non-linear radial basis kernel produced a better outcome with an accuracy of 73.3%.



However, for the cycle after applying PCA to reduced 108,000 feature columns, the SVM performance dropped to 47% of accuracy.

Furthermore, after investigating some other kernel options with 10-fold cross validation, radial basis kernel outperformed linear, poly and sigmoid kernels. Even though polynomial kernel was selected by the Grid Search CV but it only has an accuracy of 68.9%. Similarly,

tuning cost and gamma parameters did not improve the model performance thus the original setup would be kept for further modeling.

Pixel Models Summary

Comparing the outcomes given from Logistic Regression, Random Forest, and kernel SVM models, we found that both RF and SVM performs better than Logistic Regression. Also, the Kernel SVM model has a slightly higher accuracy due to smaller FP number - it does better in identifying malignant moles.

Models	Accuracy
Logistic Regression	66%
Random Forest	72.22%
Kernel SVM	73.33%

Section 4: Literature review

In the topic of skin cancer image diagnosis, many researches have suggested that shape, color and texture features were clinically important[1]. As suggested in [1], benign moles tend to be grouped in monolayers while the malignant cancerous cells are grouped in multilayer. Besides, in the Uniformity of cell size/shape, the cancerous cells are having a higher variation – parameters that represent these variations are important in determining whether the cells are cancerous or not. This motivated us to consider including such features from a clinical point of view.

From another perspective, feature extraction and representation are crucial for image analysis[2]. The most common visual features are color, texture and shape in all images. Intuitively, clinical diagnostics are closely related to these visual features when doctors are trying to identify skin cancers. Therefore, utilizing feature extraction on massive size of pixels would help construct more meaningful feature maps and help learning the diagnosis process.

For analytical point of view, statistical measures play a vital role in digital image processing[3]. Mean filters are classified as spatial filtering that capture both color and texture

[1] Classification of malignant and benign tissue with logistic regression
[2] A Review on Image Feature Extraction and Representation Techniques
[3] Importance of Statistical Measures in Digital Image Processing

features, and they are widely used for noise reduction in image processing. Standard deviation filters depict the size of variation and "dispersion" exists from the average among pixels.

Being motivated by both clinical and analytical research, we are going to focus on extracting color and texture features that represent size, shape, and color information in the original skin images:

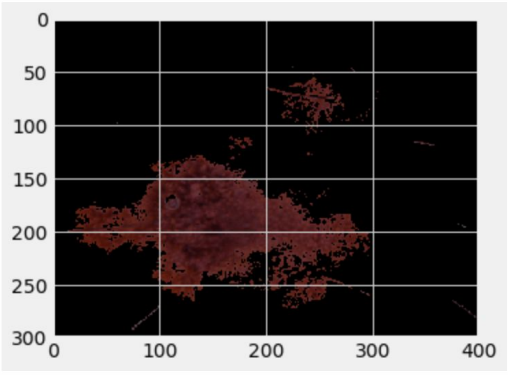
- In terms of color features, an RGB color space was specified for our project and we propose to include **Color Moments (CM)** (specifically **mean** and **standard deviation**) for the purpose of feature engineering. The RGB mean and standard variance were calculated as CM in this part of the project.
- For texture features, we incorporated **perimeters** and **areas** of marked lesion.
- We picked the **arithmetic mean** filter which brings a blurring effect to the original images, thereby reducing the local noise and variations. In addition, **standard deviation** filter was assigned to pixel color groups and the center pixel to sharpen image edge and emphasize intensity level at the edge.

Section 5: Feature engineering

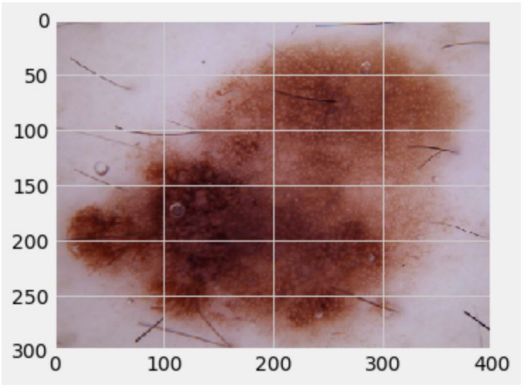
Using the pixels as image features has been one of the most common known and well-accepted ways of image analysis, however, it is hard to interpret to people who do not know the idea very well. Another downside of using pixel for image analysis is that it includes too many dimensions, in fact, each pixel contains 3 color - red, green and blue, and am image of size 300x400 contains: 300 * 400 = 120000 pixels, that is a lot.
As required by our collaborator, we decided to look for new features existing in images besides pixel that would help our analysis. Particularly, we took three different approaches to pre-process our image as well as perform feature extraction through several new approaches:

1. Hide the background of image, extract the areas and perimeter of object's contours

Often times, an image does include more than the object we are interested to analyze with, objects we are interested are considered as target and the other parts of image can be considered as the **background**. Our approach is to filter the background (healthy skin in this analysis) and take a closer look at the affected area of the skin. As shown in the two images below, after we apply a filter to remove the background of the image, it is easier for Python to catch the true contour of the affected area. We take the area and perimeter of the affected area as part of our feature.



Background Filter Applied



Before Background Filter

2. Most representative color in key area

According to research papers, color can play an important role in diagnosing skin disease. Different colors may indicate the stages or presence of skin cancer. Therefore, we performed a Kmeans clustering and selected top three most common colors in each picture, and use them as predictors to help us identify malignant or benign condition.

3. Image means and standard deviations

By considering the means of each color (red, green and blue) and standard deviations of each color, we can learn how the overall color of the graph will be, and how much changes existing in the graph, such features can be very helpful for our classification analysis.

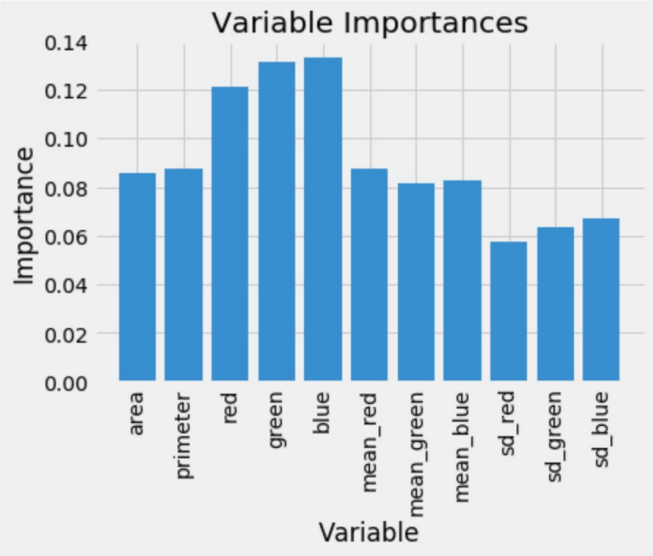
Section 6: Two Classification models with new features

After collecting new features mentioned above, we decided to build a random forest model as well as Kernel SVM model to check if the performance is improved by the new features.

The new features our model will be based on does not include the pixel, instead, it includes: 1. Area of detected contours, 2. perimeter of detected contours, 3. Color portion of key ares, 4. Standard Deviation and Mean of images. All are pretty general features which are very interpretable and easy to understand.

Random Forest Classifier using New Features

The optimal parameters of the Random Forest model, which are selected by GridSearch feature, is applied to the new feature and the model is trained through 70% of the images. This time, the accuracy of the model improved to 77.8%, compared to a pixel model, we have an increase of 5% in accuracy. However, since the dimension of the new extracted features is much less than the pixel data, the model take almost instant to run. Below is the feature importance graph generated from the RF model:



As the variable importance graph shown, all the features extracted are pretty important or they contribute a lot to classify the skin condition. The most representative color of key area is the most important feature among all.

Kernel SVM Classifier using New Features

Using the same approach as described in Question 1, the Pixel Model, we successfully fit a Kernel SVM classification model with the new features. Surprisingly, although the dimensions are way less than the pixel data, the model performance has been greatly improved.

The Kernel SVM model trained with the new features has an accuracy of 82.2% on the Test set, which is a big increase compared to 73.33% accuracy resulted from pixel data.

Classification Models Summary using New Features

In summary, after applying new feature extraction other than using pixels, we successfully build a dataset contains 11 variables (new features), the performance improved on both Random Forest Model as well the Kernel SVM model compared to a pixel model which used more than 150,000 columns (features) as shown in the table below:

Models	New Feature Accuracy	Pixel Model Accuracy
Random Forest	77.80%	72.22%
Kernel SVM	82.22%	73.33%