**Team**

Jingwei Li - jl96

Pushpit Saxena - pushpit2

Harish Chandramohan - harishc4

**Project Description**

The goal of this project is to build classification models for identifying malignant moles based on skin lesion images. The dataset contains 300 images of skin moles, of which, 150 images are of benign and 150 images are of malignant skin moles.

The first part of the project is to build 3 different classification models that are trained using the image pixel data. The image files are of inconsistent sizes, so the images were resized to a consistent format before using it to develop the models. The processing done on the images will be detailed in the later sections. Data from all 3 channels (RBG color) from the resized image will be used to develop the models.

The second part involves feature engineering to extract new features from the images, so that the features are more interpretable to the user. Research papers and existing literature in the field of skin cancer detection were reviewed to learn about clinically relevant features that are useful in detection of malignant moles. The new features will be used to train 2 new models, so that the models are more accurate, interpretable and explainable to the user (medical doctor).

**Results**

**Data Processing - Question-1 (image\_processor.py)**

The images are of inconsistent size and hence needs to be resized into a consistent format to be useful in building the model. On inspection of the images, all the images seemed to maintain a 4:3 aspect ratio with 3 channels of colors – Red, Green and Blue. The images were resized, using area interpolation method, to 133 wide x 100 high x 3 channel format to maintain the original aspect ratio and flattened to come up with 39,900 pixel-based features array for each of the 300 images. The true labels 0=benign and 1=malignant was also added to the data. The processed data was then randomly split into train and test sets, with 20% of the data from each class being set aside as test data and the remaining 80% of the data was used as training set. These data sets were then used for developing the models. The script - *image\_processor.py* was used to resize the images.

**Classification models based on Pixels**

**Literature Review**

Skin cancer is one of the most prevalent type of cancer worldwide. In the U.S, skin cancer is considered the most common type of cancer. The number of skin cancer cases has been growing over the past few decades, attributed to the higher depletion rate of Ozone layer and increase in the exposure to UV rays. Numerous studies and research have been conducted in the field of early detection of skin cancer. With the recent advancement in Machine Learning and AI, lot of research work has been conducted to use image processing and AI to detect cancer using images of skin lesions and moles.

We reviewed some of these publications to gain an understanding of the clinically relevant characteristics of the skin lesion that can be learned from the images and understanding the tools and techniques that can be used to extract these features from the images. Based on the understanding from reading few of the relevant publication the main characteristic that are useful for detecting malignant moles can be summarized based on the ABCD rule of dermatoscopy. The rule specifies the visual features associated

with malignant lesions symptoms. The ABCD acronym stands for **Asymmetry**, **Border structure**, **Color variation** and **Diameter** of lesion. These features define the basis for diagnosis of the disease and is commonly used by dermatologists.

* **Asymmetry (A)** - About half the time, a melanoma develops in an existing mole; in other cases, it arises as a new lesion that can resemble an ordinary mole. A noncancerous mole, however, is generally symmetric and circular in shape, while melanoma usually grows in an irregular, asymmetric manner.
* **Border Irregularity (B) -** Benign lesions generally have clearly defined borders. A malignant lesion, in contrast, often shows notched or indistinct borders that may signal ongoing growth and spreading of the cancer.
* **Color Variation (C)** - One of the earliest signs of cancer may be the appearance of various colors within the lesion. Because melanomas arise within pigment-forming cells, they are often varicolored lesions of tan, dark brown, or black, reflecting the production of melanin pigment at different depths within the skin.
* **Diameter (D)** - Malignant moles and lesions tend to grow larger than common moles and show typically at least a diameter of about 6mm.

**References**

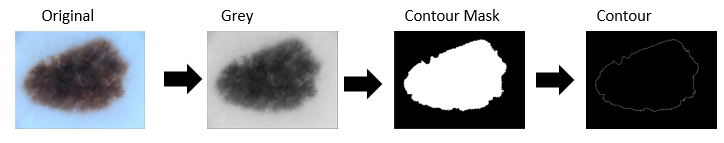
1. *American Academy of Dermatology Association (*[*https://www.aad.org/media/stats-skin-cancer*](https://www.aad.org/media/stats-skin-cancer) *)*
2. *Automatic Classification of Specific Melanocytic Lesions Using Artificial Intelligence (*[*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4739011/#B28*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4739011/#B28)*)*
3. *Feature Extraction for Skin Cancer Lesion Detection (*[*http://ijsetr.org/wp-content/uploads/2015/05/IJSETR-VOL-4-ISSUE-5-1645-1650.pdf*](http://ijsetr.org/wp-content/uploads/2015/05/IJSETR-VOL-4-ISSUE-5-1645-1650.pdf)*)*
4. *A PRELIMARY APPROACH FOR THE AUTOMATED RECOGNITION OF MALIGNANT MELANOMA (*[*https://ias-iss.org/ojs/IAS/article/view/759/662*](https://ias-iss.org/ojs/IAS/article/view/759/662)*)*
5. *Skin Cancer Diagnostics with an All-Inclusive Smartphone Application (*[*https://www.mdpi.com/2073-8994/11/6/790/htm*](https://www.mdpi.com/2073-8994/11/6/790/htm)*)*

**Feature Engineering**

Based on the understanding from the readings of existing literature, we decided to extract and use some key geometric features of the lesion as features based on the ABCD rule discussed in the previous section. The feature extraction was performed in 2 steps which involved image segmentation followed by feature extraction.

**Image segmentation**

The images were segmented using OTUS’s method. The image was first converted to grey scale and OTSU’s method uses grey level thresholding to extract the lesions from the background skin. Then lesion counter was generated using the segmented image mask. The contour was used extract the below mentioned geometric features of the image. Shown below is a sample of the intermediate images in the extraction process –



**Feature Extraction**

Asymmetry

1. Horizontal asymmetry and Vertical asymmetry: Calculated by overlapping the binary form of the warped segmented image with the mirror images in horizontal and vertical directions. The sum of all the non-zero pixels in the image is computed along the principal horizontal and vertical axis and divided by the area of the contour to get the asymmetry value. ***AS*=*Non zero pixels/Area***

Border Irregularity

1. Border Irregularity Index: The ‘Border irregularity’ feature is generally defined as the level of deviation from a perfect circle and measured by the irregularity index using the below formula – **BI\_index = *Perimeter of contour^*2 /(4\**π\*Area of contour)***

Diameter based features

1. Horizontal diameter: This is the horizontal diameter along the horizonal principal axis of the image contour
2. Vertical diameter: This is the vertical diameter along the vertical principal axis of the image contour
3. Area of lesion: This is the area of the lesion calculated by counting the non-zero pixels in the contour mask shown above
4. Perimeter of lesion: This is arc length of the image contour

**Classification model based on new Features**