**Group Project: Skin Cancer Diagnostics**

* Data: 300 Images of moles can be downloaded from [here](https://www.dropbox.com/s/15va32gdq42ukwh/Skin%20Cancer%20Data.rar?dl=0). The data contains 150 benign and 150 malignant skin images, separated into two different folders.
* **Question 1**: As a data scientist, you are asked to construct classification models for identifying malignant moles based on the pixels (RGB colors) of these images.
* **Question 2**: Your collaborator does not like the pixel model because it is not interpretable. Can you propose new data processing/feature engineering approaches that improve the interpretability and model accuracy?

**Detailed requirements**

* [Submission Guideline]
  + Pick **one team member as a team lead to submit the full report and supplementary file** to his/her submission portal (coursera or compass2g). In this report, you should clearly indicate all of your team members (their names and NetID) on the first page.
  + The full report should be in .pdf format and **no more than 10 pages**.
  + All other team members will submit a one-page pdf file, only to indicate the information of your team lead, and also list all team members.
  + **DO NOT** have multiple team members summitting the same/different final reports. Doing so will lead to a 5-points penalty to all of your team members, and if those reports are different, the lowest one will be used.
  + All members should name their reports the same way as homework, e.g., Project2\_yourNetID.pdf.
* Grading. Your report should contain the following:
  + [10 Points, 1 page] Project description and summary. This part should summerise your goal, your approach, and your results for both questions.

Skin cancer, including melanoma and nonmelanoma skin cancer, is one of the most prevalent carcinomas worldwide. Clinal confirmatory test of skin cancer often requires procedure such as a biopsy, which is invasive and time-consuming. Building reliable models that can classify skin cancer based on dermoscopic images may extensively accelerate the diagnosis process and prevent the suffering of a biopsy. Before training models, images were preprocessed with cropping, resizing, blurring, and dilation to eliminate irrelevant objects such as black frames, hair, bubble, and ruler on the lens. [Model training part]

With the goal of enhancing interpretability of the models, skin lesion was automatically identified by border detection and 22 features were engineered to reflect skin lesion’s shape characteristics (perimeter, area, border irregularity, shape compactness) and color characteristics (pixel mean and standard deviation of normalized RGB, HSV, CIE L\*a\*b color channels). [Model training part]

* + [5 Points, half a page] Data processing for Question 1. Describe how you process the data so that it can be analyzed to answer question 1.

The raw data are consisted of 300 color photos of different sizes. As the photos were taken with a dermoscope on skin with a variety of conditions, the goal of photo preprocessing is to eliminate irrelevant objects in the view, such as the black frame, the dark area beyond the ocular lens, the hair, the bubbles, and the ruler on the object lens. To start with, black frames and dark areas were identified by converting photos to grayscale and checking the pixels in each row/column. A row was considered a black frame when over 50% of the pixels were black (normalized pixel value < 0.1) while a column was considered intersecting a dark area when over 20% of the pixels were black (Celebi et al., 2008). These rows and columns were cropped from the original photo. Next, all the photos were further cropped and resized to 300 x 225 pixels with linear interpolation to reduce computation time. To remove hair and the ruler, image dilation was conducted with a size-3 square structuring element (Do Hyun Chung & Sapiro, 2000). Furthermore, bubbles were blurred by applying a 5 x 5 median filter (Masood & Al-Jumaily, 2013). Eventually, each photo was converted into a vector of 202,500 pixels, with pixels in R, G, B channels concatenating one another sequentially.

* + [30 Points, within 5 pages] Classification models based on pixels.
    - For this question, you must choose **three** different classification models (10 points for each model) to identify malignant moles. Please note that your models have to be significantly different, e.g., Lasso and Ridge are treated as the same type of model. Besides, you **cannot use any deep learning model**.
    - To evaluate your models, you should consider appropriate tunings and validation.
    - Demonstrate your results in an intuitive way
  + [40 Points, within 4 pages] In this question, we need to consider new feature engineering approaches to make the model more interpretable and accuracy. The goal is to convince your collaborator (a medical doctor) that the model may suggest some interesting characteristics for identifying malignant moles. For example (not necessarily good suggestion), you could consider the overall color, the color variation, the shape, etc. To perform the following sub-questions:
    - [10 Points, 1 page] Literature review. You should search and read existing literature and summarize clinically relevant characteristics that could be used for skin cancer image diagnosis. There is no limitation on what type of literature you could use. However, the goal should be motivating your feature engineering approaches from a clinical and analytic point of view. Please give appropriate citations to the literature you read.

Among all the clinical diagnostic criteria proposed in the literature, the “ABCD” criteria stand out as the most widely-adopted method. The acronym stands for asymmetry, border irregularity, color variegation, and diameter > 6 mm. Alternative criteria like the Glasgow 7-point checklist also mentions relevant characteristics such as varying size, shape, color (blue-white structure) and the observation of bleeding and inflammation. (Rigel, Russak, & Friedman, 2010). It has also been reported that features such as vascular growth, thickness of the lesion, evolving lesion size are associated with higher risk of developing skin cancer (Martinez & Otley, 2001). Therefore, the shape and color of skin cancer may be crucial to classification.

* + - [10 Points, 1 page] Feature engineering. Motivated by what you have read (or your understanding), process the data in a reasonable way such that the new variables are more intuitive to your collaborator/clinicians. You need to describe clearly what is your data processing criteria and how your variables are calculated.

Based on the literature review, several shape and color features have been engineered with goal of capturing the essence of each skin cancer image.

Before developing features, the skin lesion or the region of interest (ROI) was identified and the border of skin lesion was detected. Specifically, each image was converted to a grayscale image, which was then binarized by an automatically determined threshold. A grayscale pixel was converted to 1 if it exceeded the threshold and 0 if otherwise. This resulted in a black-and-white image where the black region represented the ROI and the white represented the background.

Shape features such as perimeter and area were first engineered. Perimeter (P) of an ROI was defined as the number of pixels on the border and area (A) was defined as the number of pixels covering the entire ROI. On top of these 2 basic shape features, irregularity index () and compactness (the ratio between the area of the ROI and the area of a circle with the same perimeter as the ROI) were constructed (Celebi et al., 2007; Golston, Stoecker, Moss, & Dhillon, 1992).

Color features such as the mean and standard deviation of pixels in each color channel were constructed. Generally, each image was split by its 3 color channels and the pixels within the ROI were extracted and subjected to calculation of the mean and the standard deviation. To create color features possessing such characteristics as invariance to illumination intensity, perceptual uniformity, and decoupling of chrominance and luminance, we have chosen 3 different color spaces (normalized RGB, HSV, CIE L\*a\*b) to calculate the means and standard deviations (Celebi et al., 2007; Green, Martin, Pfitzner, O’Rourke, & Knight, 1994).

In summary, a total of 22 features were engineered for each image including 4 shape features and 18 color features.

Asymmetry index: (Celebi et al., 2007)

* + - [20 Points, 2 page] Classification models based on new features. Fit **two** different classification models to identify malignant moles. You can either use the ones from Question 1 or use some new models if you believe they may perform better on the new features. Same requirements of Question 1 apply to this part. Besides, you should focus more on variable selection and interpretation.
  + [15 Points] General organization, neatness, readability, and use of R. Is your report easy to read with clear logic? Is it written in a manner such that a reader does not need to be very familiar with the data? Are plots, tables, etc. informative and properly displayed? Are you properly utilizing R markdown to have a clean report? Is irrelevant code/output hidden?

## Reference

Celebi, M. E., Kingravi, H. A., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., Moss, R. H., … Menzies, S. W. (2008). Border detection in dermoscopy images using statistical region merging. *Skin Research and Technology : Official Journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)*, *14*(3), 347–353. https://doi.org/10.1111/j.1600-0846.2008.00301.x

Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, *31*(6), 362–373. https://doi.org/10.1016/j.compmedimag.2007.01.003

Do Hyun Chung, & Sapiro, G. (2000). Segmenting skin lesions with partial-differential-equations-based image processing algorithms. *IEEE Transactions on Medical Imaging*, *19*(7), 763–767. https://doi.org/10.1109/42.875204

Golston, J. E., Stoecker, W. V., Moss, R. H., & Dhillon, I. P. S. (1992). Automatic detection of irregular borders in melanoma and other skin tumors. *Computerized Medical Imaging and Graphics*, *16*(3), 199–203. https://doi.org/10.1016/0895-6111(92)90074-J

Green, A., Martin, N., Pfitzner, J., O’Rourke, M., & Knight, N. (1994). Computer image analysis in the diagnosis of melanoma. *Journal of the American Academy of Dermatology*, *31*(6), 958–964. https://doi.org/10.1016/S0190-9622(94)70264-0

Martinez, J.-C., & Otley, C. C. (2001). The Management of Melanoma and Nonmelanoma Skin Cancer: A Review for the Primary Care Physician. *Mayo Clinic Proceedings*, *76*(12), 1253–1265. https://doi.org/10.4065/76.12.1253

Masood, A., & Al-Jumaily, A. A. (2013). Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms. *International Journal of Biomedical Imaging*, *2013*. https://doi.org/10.1155/2013/323268

Rigel, D. S., Russak, J., & Friedman, R. (2010). The evolution of melanoma diagnosis: 25 years beyond the ABCDs. *CA: A Cancer Journal for Clinicians*, *60*(5), 301–316. https://doi.org/10.3322/caac.20074