Data Selection Proposal (Gargi Singh and Jehan Dastoor)

The Dataset (https://challenge2018.isic-archive.com/task3/)

We chose this dataset because it is robust, with lots of input image data as well as relevant patient data (summarised in the metadata.csv) which can be used to optimize the most accurate predictive model. The dataset is also segmented into training, validation, and testing input data, which will make the training/validation/testing process more straightforward and convenient. Moreover, the distribution of the lesion types present has been crafted to reflect "real world" settings, in which there are more benign than malignant lesions but malignancies are overrepresented. Training the model on these data will cause it to be more in tune with real world conditions and subsequently more generalizable.

Methodology

a. The dataset has a "metadata.csv" file that identifies which image belongs to each row (with a tag).
    i. The text labels can be converted into integers, for example diagnosis type bkl to 0 etc. The dataset has already been edited to not have empty rows so nans are not an issue.
    ii. Image transformations will be needed to transform the image data into pixels or RGB values that can be used by our ML model.
b. ML Model:
    i. We aim to predict the type of skin lesion from a user provided image.
    ii. A convolutional neural network will be used for feature detection of types of skin lesions. GoogleNet is one option we have in mind. (possibility of a capsule net)
    iii. Pros: Automatic detection of important features. Cons: Need a large dataset for training (which we have), the orientation or position of the object is not recorded.
c. Evaluation Metric: According to the task outlined, we would be using a normalized multi-class accuracy metric (balanced across categories). If this does not work, we aim to use an AUC (area under the receiver operating characteristic curve) metric.
d. Final Conceptualization: The user would input the age, localization and sex of the patient whose image was taken in textboxes/drop-down lists and (most importantly) upload their image. Our model would then process the image and output a diagnosis (and possibly a methodology for diagnosis). The webapp will also display the percentage confidence in the answer which we will aim to have above 97% as this is the typical confidence that is accepted in the medical industry (before a doctor uses the suggested diagnostic methodology).

Application

The aim is to produce a React.js with flask webapp. If time permits we might also try to deploy the app with Herokuapp, otherwise we will just run it locally. The web UI can be designed using Figma and converted to code through the Anima extension after some manipulation.