Capstone 3 - Proposal

By: Rayees Ahamed - Springboard Data Science Career Track Student

Title: Cell Images Classification - Infected (Parasitized) or Uninfected

Problem statement:

Predicting the parasite infected and uninfected cells from microscopic images

Context:

Infectious diseases such as malaria, dengue are caused by parasites like viruses. After infection, the patient's cells start spreading these parasites all over the body. Once the body cells are infected with parasites, that is called **'parasitized' or 'infected'** cells. Other hand, the cells that are not infected with parasites are called **'uninfected'**. So, it is crucial to detect these infected (parasitized) cells before they cause serious illness to patients. We can identify these infected cells using their images which are stained with chemical dyes. Generally, two types of dyes are used to identify the 'infected' and 'uninfected' cells: 1. eosin **(pink)** and 2. hematoxylin **(blue)**. Simply to understand, the images which contain the blue stains are infected with parasites, whereas cell images containing only pink are classified as uninfected cells. Here, we aim to classify the infected and uninfected images by training more than 10000 images that will help in predicting the right class upon testing.

Criteria for success:

1. Labeling, feature extraction and training the two classes of images
2. Building and testing a reliable model that can predict the infected images with high accuracy and precision

Scope of solution space:

The scope of the project lies in preliminary image processing and labeling, followed by extracting the right features from infected and uninfected image data. Selecting the appropriate classification model such as CNNs to train the image sets. Choosing the optimum model evaluation metrics (accuracy and precision etc.) and deploying the right model with highest performance scores.

Constraints within solution space:

Infected image classes may have several features, so selecting the right features may be a tedious process. Also, the training phase may require an extensive amount of time to complete due to the large number of images. Final model can be used only for research purposes and cannot be used for clinical treatments.

Stakeholders to provide key insight:

1. Dipanjan Sarkar - Data Scientist & Mentor at Springboard
2. Rayees Ahamed B - Springboard Data Science Student

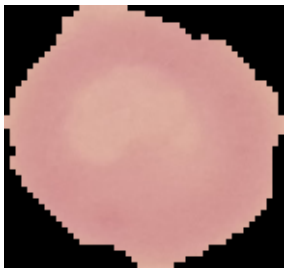
Key data sources:

- Kaggle Data Repo - https://www.kaggle.com/brsdincer/cell-images-parasitized-or-not

Data can be downloaded from the above linked page. Originally deposited by Baris Dincer in Kaggle datasets page. It contains more than 10000 images to train and test to identify the infected and uninfected class of images.

Sample images:

Infected



Uninfected

Project deliverables:

- Github repository with code notebooks
- Project report in a PDF format
- Slide deck containing each step summary