Capstone Project
Malaria Detection
Milestone 1

Joon Jung

Problem Definition:

1. The context - Why is this problem important to solve?
2. The objectives - What is the intended goal?
3. The key questions - What are the key questions that need to be answered?
4. The problem formulation - What is it that we are trying to solve using data science?

Malaria is a contagious disease that causes more than 229 million malaria cases and 400,000 malaria-related deaths reported over the world in 2019. Almost 50% of the world's population is in danger from malaria and children under 5 years of age are the most vulnerable population group affected by malaria as they accounted for 67% of all malaria deaths worldwide. Any contribution of reducing the numbers of malaria-related deaths or streamlining diagnosing process can benefit millions of people and it is a very important problem that we must face and solve.

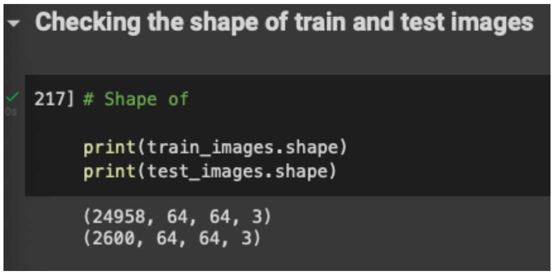
The pathology of malaria is that it is a contagious disease caused by Plasmodium parasites that are transmitted to humans through the bites of infected female Anopheles mosquitoes. The parasite enters the blood and begin damaging red blood cells (RBCs) that carry oxygen, which can result in respiratory distress and other complications. The lethal parasites can stay alive for more than a year in a person's body without showing any symptoms, which means that late treatment can cause complications and could even be fatal. This leads to the fact that early, fast, and accurate diagnosis of malaria is very crucial. However, the problem rises where the traditional diagnosis of malaria in the laboratory requires careful inspection by an experienced professional to discriminate between healthy and infected red blood cells. It is a tedious, time-consuming process, and the diagnostic accuracy can be adversely impacted by inter-observer variability since it heavily depends on human expertise.

This is where a data science can provide a solution. An automated system using automated classification techniques using Machine Learning (ML) and Artificial Intelligence (AI) have consistently shown higher accuracy than manual classification by human, and it can drastically help with the early and accurate detection of malaria. It would be beneficial to propose an efficient computer vision model that performs malaria detection using Deep Learning Algorithms, where it can identify whether the image of a red blood cell is that of one infected with malaria or not, and classify the same as parasitized or uninfected.

Data Exploration

1. Data Description - What is the background of this data? What does it contain?
2. Observations & Insights - What are some key patterns in the data? What does it mean for the problem formulation? Are there any data treatments or pre-processing required?

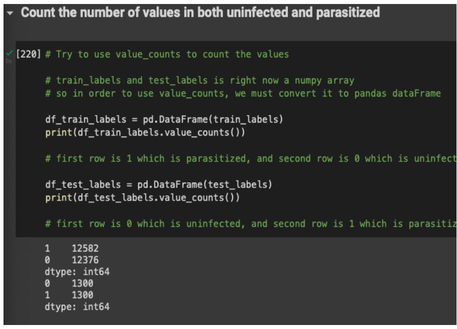
The data contains colored images of red blood cells that contain parasitized and uninfected instances, where parasitized cells contain the plasmodium parasite, and uninfected cells are free of plasmodium parasites but could contain other impurities.



Checking the shape of train and test images

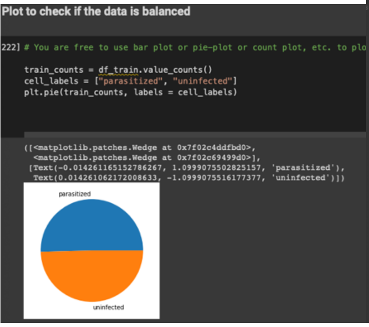
```
217] # Shape of

     print(train_images.shape)
     print(test_images.shape)

     (24958, 64, 64, 3)
     (2600, 64, 64, 3)
```

There are 24958 images in training data set and there are 2600 images in test data set.



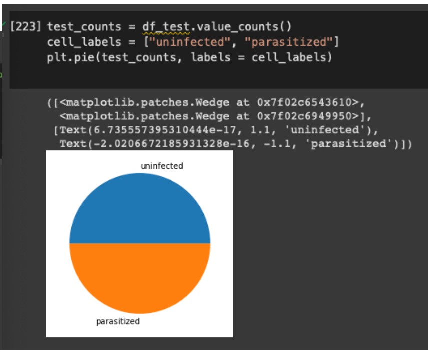
Count the number of values in both uninfected and parasitized

```
[220] # Try to use value_counts to count the values

      # train_labels and test_labels is right now a numpy array
      # so in order to use value_counts, we must convert it to pandas dataFrame

      df_train_labels = pd.DataFrame(train_labels)
      print(df_train_labels.value_counts())

      # first row is 1 which is parasitized, and second row is 0 which is uninfect

      df_test_labels = pd.DataFrame(test_labels)
      print(df_test_labels.value_counts())

      # first row is 0 which is uninfected, and second row is 1 which is parasitiz

      1    12582
      0    12376
      dtype: int64
      0    1300
      1    1300
      dtype: int64
```



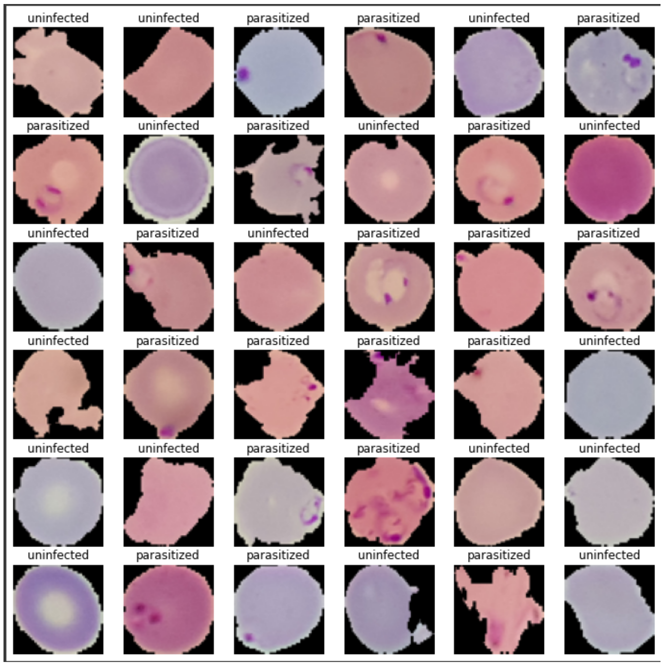
Plot to check if the data is balanced

```
222] # You are free to use bar plot or pie-plot or count plot, etc. to plo

     train_counts = df_train.value_counts()
     cell_labels = ["parasitized", "uninfected"]
     plt.pie(train_counts, labels = cell_labels)

     ([<matplotlib.patches.Wedge at 0x7f02c4ddfbd0>,
       <matplotlib.patches.Wedge at 0x7f02c69499d0>],
      [Text(-0.014261165152786267, 1.0999075502825157, 'parasitized'),
       Text(0.014261062172008633, -1.0999075516177377, 'uninfected')])
```



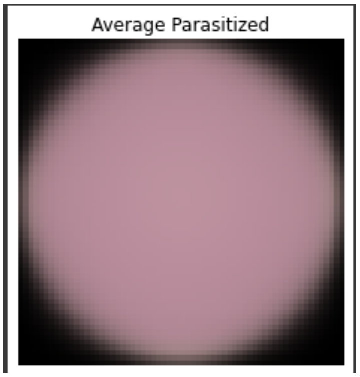
```
[223] test_counts = df_test.value_counts()
      cell_labels = ["uninfected", "parasitized"]
      plt.pie(test_counts, labels = cell_labels)

      ([<matplotlib.patches.Wedge at 0x7f02c6543610>,
        <matplotlib.patches.Wedge at 0x7f02c6949950>],
       [Text(6.735557395310444e-17, 1.1, 'uninfected'),
        Text(-2.0206672185931328e-16, -1.1, 'parasitized')])
```

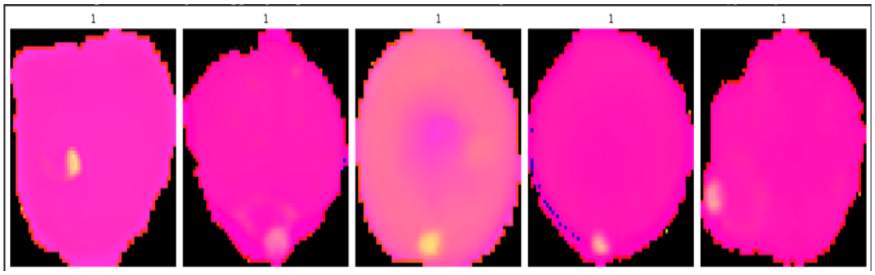
There are 12582 parasitized cell pics and 12376 uninfected cell pics in the training data, and there are 1300 pics for each parasitized and uninfected cell groups in test data set, which can also be seen in the pie chart.

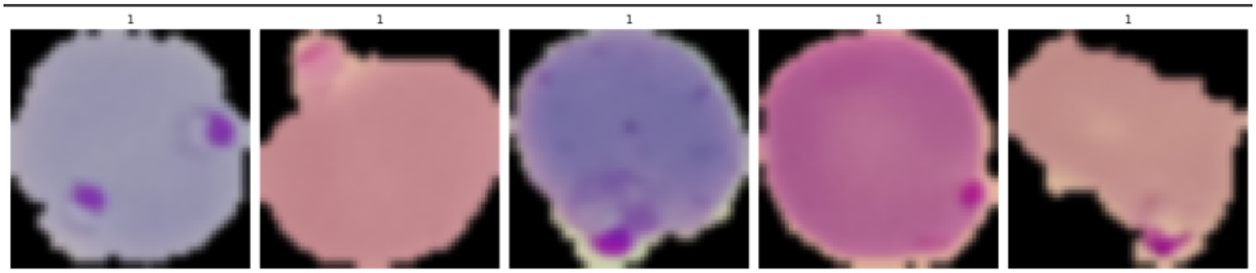


This is how the pics of the red blood cells look like. They are labeled as parasitized or uninfected. It is evident that presence of Plasmodium parasite seems to show visually on pictures as blue marks in the cell, and the cell shapes tend to be more crippled on parasitized cells than uninfected cells, although that needs to be further evaluated.

Various data transformation techniques, such as average mean image, converting RGB image to HSV, and applying Gaussian Blurring, were used on the image data to evaluate further direction and options.







Proposed Approach

1. Potential techniques - What different techniques should be explored?
2. Overall solution design - What is the potential solution design?
3. Measures of success - What are the key measures of success to compare potential techniques?

Since we are approaching image data set, using Convolutional Neural Networks (CNNs) should be explored. The presence of blue mark can be anywhere within the cell and difference in cell shapes would also contribute to making distinction between parasitized cell and uninfected cell, meaning we need advanced local spatiality and better detection of edge, curve, color, etc that can be achieved with use of CNNs. Using ANN would flatten the current image array of 4D array into 1D array, which would not be helpful with local spatiality and detecting patterns in different image data. We can also apply convolutional layers and filters upon the image data. The key measures of success to this proposed method would be the accuracy of the model %wise compared to the validation data set. Confusion matrix that shows how much of errors and what type of errors that model would also be used as validating our model and determining if it was successful or not.