MEDICHAL: Detection of malaria using cells images

C.Veyssiere, T.Deschamps Berger, N.Devatine,
R.Hamidi, S.Monteiro and X.Xu

September 2019

Subject: Malaria

The malaria is a disease that is spread by a bite of an infected female mosquito. The malaria is caused by the Plasmodium which is a genus of parasites. We know five types of Plasmodium that can infect humans. The malaria is a global concern, every 2 minutes, a child dies of malaria. And each year, more than 200 million new cases of the disease are reported, mostly in Africa. The early diagnosis could help treat and control the disease.

Dataset

For this challenge we decided to use a dataset from the National Institutes of Health (NIH) which is the primary agency of the United States government responsible for biomedical and public health research. This dataset includes a total of 27,558 cell images with equal instances of parasitized and uninfected cells.

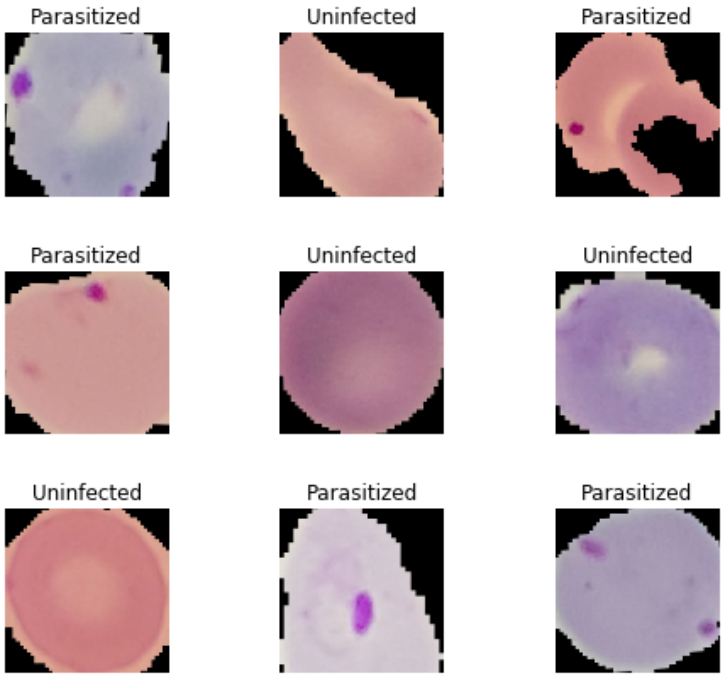
https://lhncbc.nlm.nih.gov/publication/pub9932

Since the data are images we will first have to preprocess them in order to obtain an array of practical data.

Analysis

Once the human get infected by the malaria, the Plasmodium parasite begin to spread and appears in the host's blood. By detecting this Plasmodium through a blood smear test and the "thick drop" method, it is possible to know whether a human has contracted the malaria.

Example of parasitized and uninfected cells



As we can see, the images of the cells vary from one to another. They differ in size, shape and colour.

Method

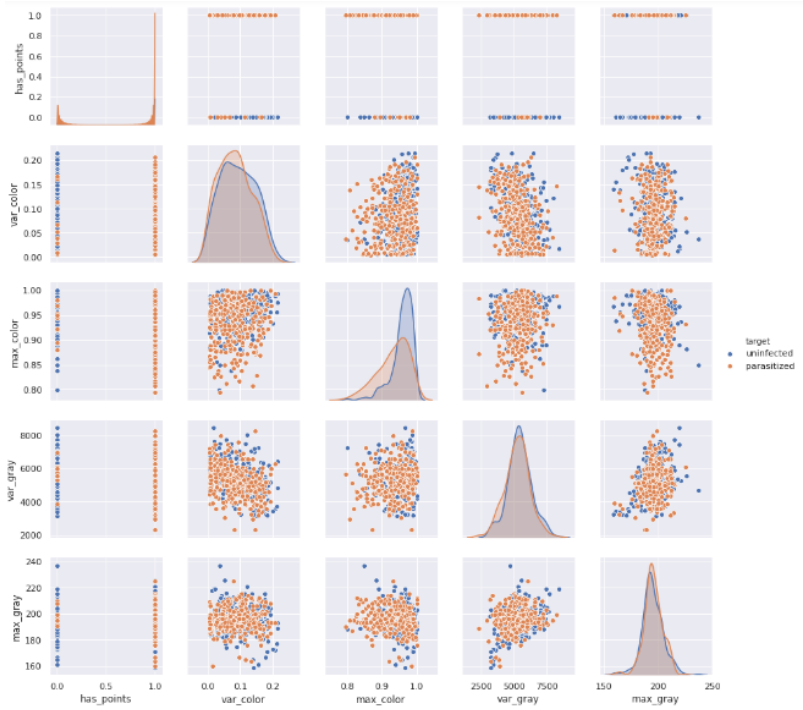
Our goal will be to detect cells that contain malaria, as a consequence, this is a problem of binary classification. Thus, we already identified some tasks that we will have to perform.

Images pre-processing. The data-set contains a total of 27,558 images, however, they don't have the same size. So, to transform them to a table, we need to find a suitable size and resize them.

Images augmentation. We don't want the result to be affected by the flips, translations, rotations of the objects in the picture. And the images perhaps aren't enough to train the model, such as large neural network, So we will use *images augmentation* to improve the robustness of our model.

Features creation. Our first approach was to created basics features. For each images, we computed the variance and the maximum of red pixels ("var_color" and "max_color"). Then after a transformation of our images in gray-scale, we computed in the same way the variance and the maximum of the pixels color ("var_gray" and "max_gray"). Finally we created a last feature called "has_points" to determine if the image has some pixels that represent the Plasmodium in our cells. To do that, we computed the mean of our pixels color for each image. And created two thresholds to filter the pixels. The first threshold filter the pixels below our mean minus a parameter (to dispose of the pixels that represent the blood). And the second threshold filter the pixels above our mean plus a parameter (to dispose of the black pixels in the background). Thus the pixels gathered should represent the Plasmodium. For each image the feature "has_point" return 1 if we have pixels and 0 otherwise.

Data visualization. We also tried to identify patterns with a pairplot, which could give us structures for each class depending on 2 features.



Preliminary results

For this project we already have baseline results from Kaggle obtained using deep learning algorithms with an accuracy of 97%. This challenge offers the possibility to obtain very good results by using the good algorithms and the good features.

First prediction We first tried to run a Support Vector Classification with our features on the variance and the maximum of red color on our test images data-set. Our first result was 55% of success which was not a good result, it's just a little bit better than the random classification.

Second prediction We decided to create more features (cf Method part) on our images in gray-scale. As for now with a Random forest and all our created features we are able to give a prediction of 83% of the infected cell or not in our patients.