Milestone Report 1

Malaria Detection: Medical Image Analysis

Credit for the picture: askideas.com

Ozkan Serttas

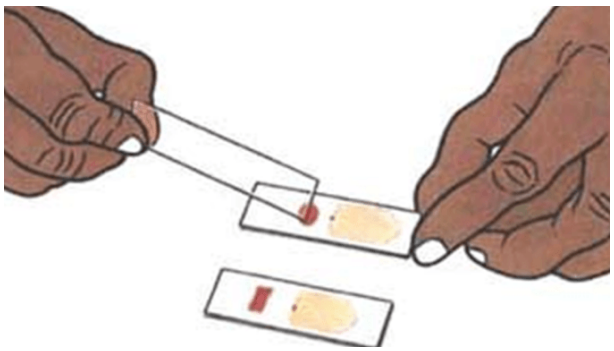
July 2018 Cohort

Milestone Report 1

Medical Image Analysis

Motivation

Malaria cases must be recognized promptly in order to treat the patient in time and to prevent further spread of infection in the neighborhood via local mosquitoes. Malaria is considered a potential medical emergency and should be treated accordingly. Delay in diagnosis and treatment is a leading cause of death in malaria patients. Approximately 400, 000 deaths every year world wide which is mostly affecting poor areas. The way malaria is diagnosed is a bit tricky as it requires both clinical diagnosis and microscopic diagnosis. Clinical diagnosis phase involves checking patient's symptoms such as fever, chills, sweats, muscle pains, etc. and examining physical conditions like tiredness, perspiration and so on. Microscopic diagnosis is examining patient's blood sample under the microscope to identify parasites.



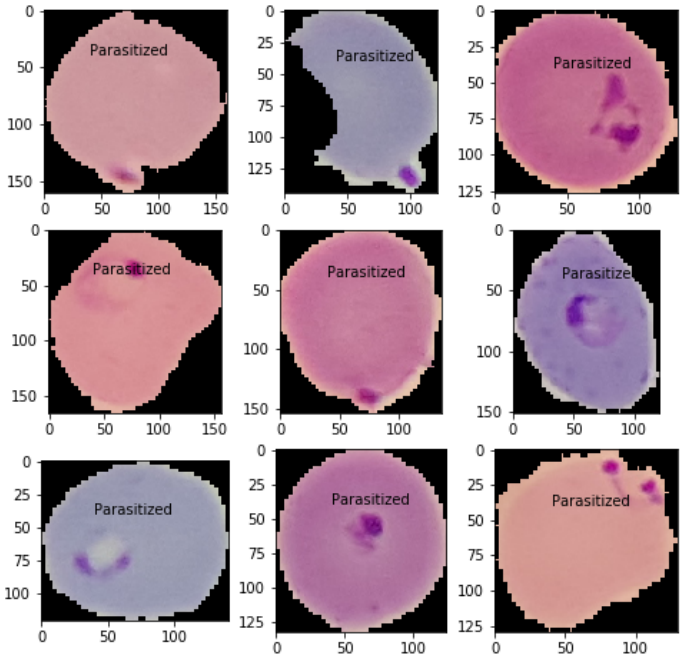
Preparation of Thick and Thin Blood Smear: microbeonline.com

Building up a model to help diagnosing malaria cases can help automating the process and clinicians so that they can spend more time on treating patients.
In this project we will explore a few image classification algorithms available to identify the best performing model for this job.

Description of The Dataset

The dataset used in this project is obtained from the U.S. National Library of Medicine, which is set of blood smear images collected by researchers. The images were manually annotated by experts at the Mahidol-Oxford Tropical Medicine Research Unit in Bangkok, Thailand. The dataset contains a total of 27,558 cell images with almost equal instances of parasitized and uninfected cells.

In order to feed image data to machine learning models we use a few preprocessing methods such as scaling, resizing, transforming, rotating and shifting. The figure below show raw image samples before preprocessing.



Data Preprocessing

Dataset is downloaded from NIH website in csv form which includes labeled cell images. We first write a source code to divide dataset into three sets those are training, validation and test sets. Train set contains 80 % of the whole data while test and validation sets are sharing 10 % respectively. After splitting dataset, we seek for the best size to resize images so that our algorithm would work at optimum pace. Average pixel dimension was obtained as [133.16, 132.61, 3], while median pixel dimension was [130, 130, 3]. However, since we ran our models on CPU, we picked even smaller dimensions as such [64, 64, 3] to save time on modeling phase.

	SampleSize	Height	Width	ColorChannel
TrainData	22235	64	64	3
Validation	2470	64	64	3
Test	6177	64	64	3

The figure below illustrates the blood smear images after resizing process. We will also apply transformation, rotation and shifting techniques later in modeling section using Keras data augmentation.