Computer Vision for Malaria Detection - Milestone 1

Sulian Thual, MIT ADSP October 2021

We build a computer vision model for malaria detection, as part of a capstone project for the MIT Applied Data Science Program (ADSP) from August-November 2021. In this first report (Milestone 1), we determine the problem definition, initial data exploration and proposed approach.

1 Problem Definition

1.1 Context

Malaria is a disease caused by parasites (plasmodiums) transmitted through mosquito bites and that infect the red blood cells. Typical symptoms of malaria include fever, fatigue, headaches, and, in severe cases, seizures and coma, leading to death. As shown in Figure 1, malaria is common in tropical and subtropical countries. In 2016 for example there were about 214 million cases of malaria globally and about 438,000 malaria deaths. Most deaths occur among children under 5 in Africa, where a child dies almost every minute from malaria. Malaria also has a significant negative effect on economic development. In Africa for example, it is estimated to result in losses of US$12 billion a year due to increased healthcare costs, lost ability to work, and adverse effects on tourism.

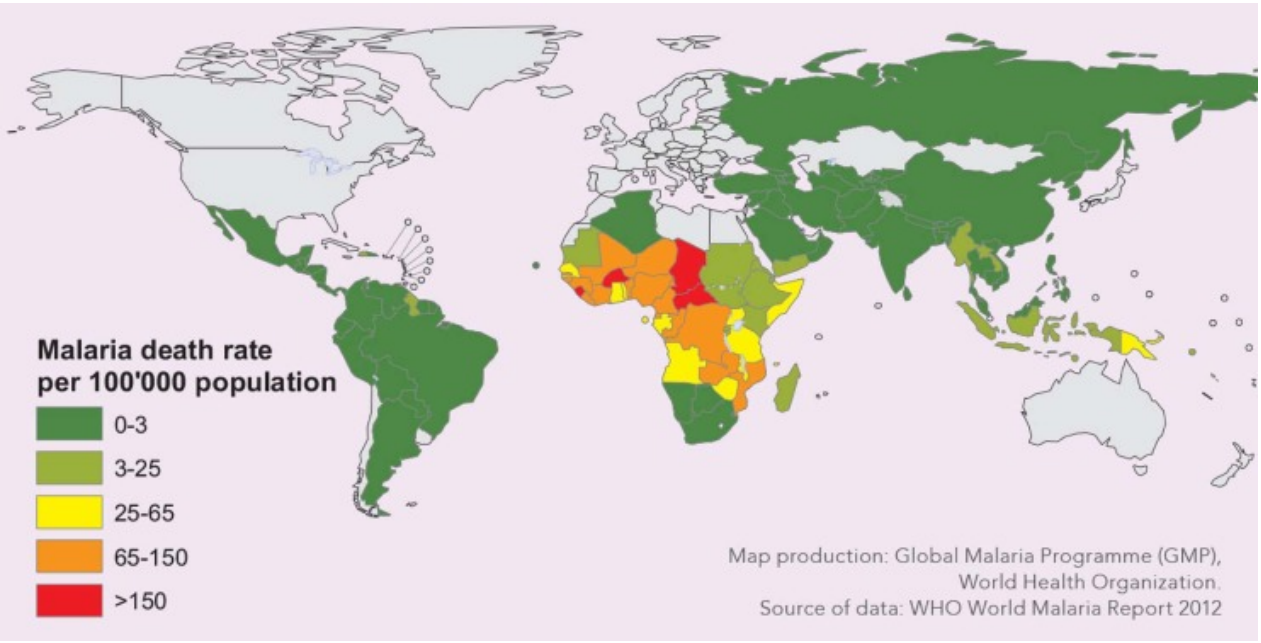

Figure 1: Worldwide malaria death rates from WHO World Malaria report 2012. Adapted from Poostchi et al., 2018.

1.2 Objectives

The goal of this project is to develop a computer vision model (a common machine learning method) for malaria diagnostic from red cell images. Hundreds of millions of blood films are examined every year for malaria. Accurate parasite counts are essential for malaria diagnosis, measuring drug-resistance and drug-effectiveness as well as classifying disease severity. Manual diagnosis is the most common method, but it is labor intensive and depends heavily on the skill of the microscopist. Automated diagnosis, while still in early development, has the advantage of providing a more reliable and standardized interpretation while also reducing workload and costs and potentially scaling to more patients.

1.3 Key Questions

In order to build our predictive model for malaria detection, we first need to understand the important characteristics of red cells to focus on (malaria shape and life cycle, etc), and whether can they succesfully be encoded by our computer vision model. Second, we must determine suitable model characteristics (e.g. architecture, training method and measures for success). Finally, we should consider how our model could be used in practical settings for diagnosis.

1.4 Problem Formulation

Detecting malaria from red cell images is an image classification problem. Let X (called feature vector) be a vector concatenating input red cell images, and Y (called label) the class of each image (uninfected or parasitized by malaria). By convention, we will refer to $Y = 1$ (parasitized) as the positive class. Our goal is to build a model with parameters θ for the relationship $Y = F(X; \theta)$. We will train the model to learn the parameters θ on a training dataset then determine its predictive power on a separate test dataset.

2 Data Exploration

2.1 Data Description

In order to build our computer vision model, we analyze a dataset of 24958 train and 2600 test images (colored) taken from microscopic images. It is obtained from the official NIH Website (and is also featured on kaggle). Each image frames and crops a single red cell taken from a thin blood smear. All images have been resized to 64x64 pixels with 3 RGB values (raning from 0 to 255), meaning the feature vector X is of size 64x64x3 for each sample. The images are either labelled as uninfected or parasitized, taking label values $Y = 0$ or $Y = 1$ for each sample, respectively. Both train and test datasets have a 50% split between uninfected and parasitized.

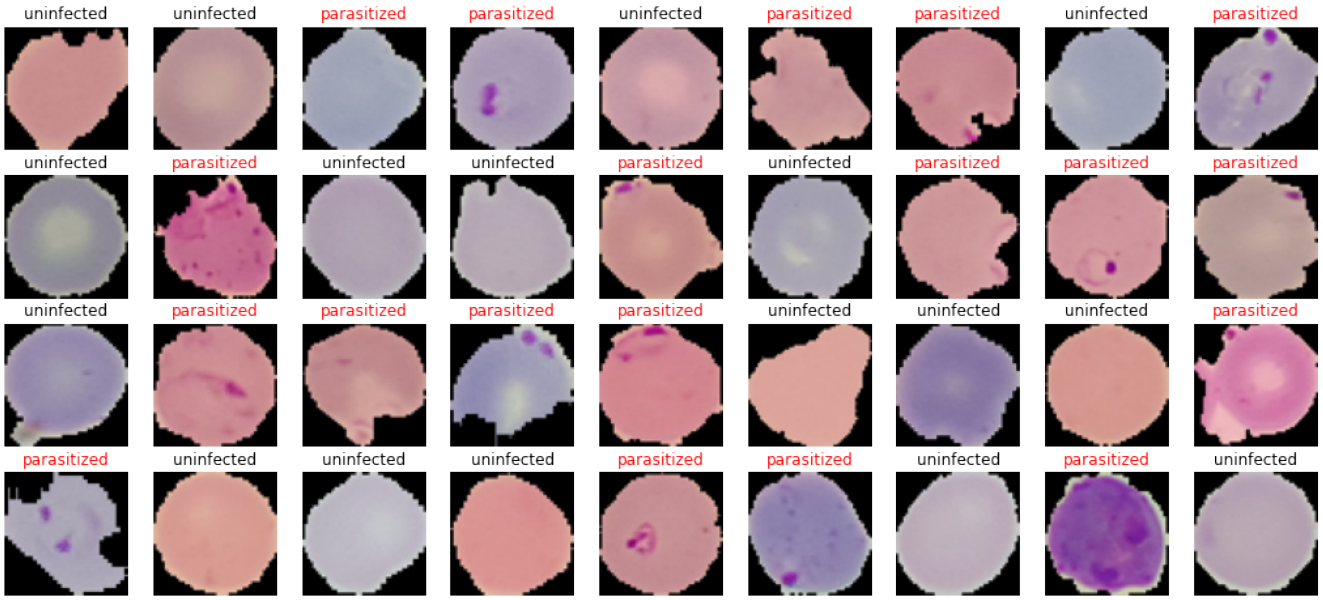

Figure 2: Examples of red cell images from the database (uninfected or parasitized).

2.2 Observations and Insight

Figure 2 shows examples of images taken from the database. Cell color may vary from light pink to light blue regardless of the label. Uninfected cells are uniform with sometimes a lighter center (which is due to the "donut shape" of red cells). Parazited cells in contrats contain or several malaria parasites. The parasite can take various shapes (e.g. rings, elongated, crescents, clefts) which depends on its type (there are 4 malaria types) and the stage of its life cycle (see e.g. the official CDC DPDx website for illustrations). Either uninfected or parasitized cells can look deformed, which is an artifact of the framing and cropping. In reality, both types of cells should be ellipsoidal (except for parasitized cells at the Maurer's clefts life-cycle stage).

Figure 3 shows examples of filters applied to a parasitized cell. These filters provide alternative visualizations that are very useful for understanding. However, filters result in data loss therefore it is unlikely that filtered images should be used to train our computer vision model.

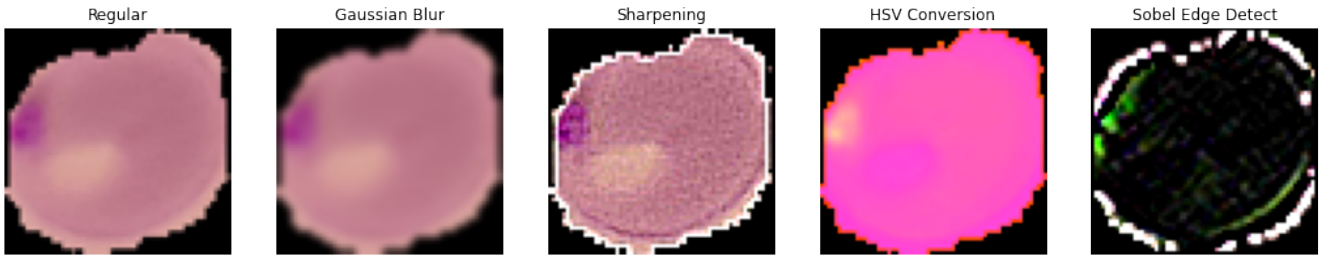

Figure 3: Examples of image filters on a single paratisized cell.

3 Proposed Approach

3.1 Potential techniques

The detection of malaria from red cells images is an image classification problem. Although many Machine Learning methods could be used to solve it, the most suitable model to use as our computer vision model is a Convolutional Neural Network (CNN). CNNs are a class of artifical neural network most commonly applied to analyze images, and differ from regular Artifical Neural Networks by their use of convolution filters. A major advantage of CNNs is that they use relatively little pre-processing compared to other image classification algorithms, meaning that they optimize the filters (or kernels) through automated learning.

3.2 Overall Solution Design

Figure 4 shows a sketch of the CNN model to use for malaria detection. Its architecture consists of several convolutional layers for feature extraction, followed by fully-connected layers (and a final sigmoid activation function) for binary classification. We will train the model on the train dataset in order to learn its parameters (weights). Once trained, we will assess its performance on the test dataset. When deployed, for any given input image of a red cell the CNN will predict if it is uninfected or parasitized.

While these are the general features of the CNN, many of its characteristics will have to be determined through experimentation. This includes the architecture (types of layers, regularization), training method (initialization, type of gradient descent method, transfer learning and/or data augmentation, etc) as well as measures for success. Typically, we will repeat cross-validation as well as hyperparameter search depending in order to improve predictions. Advanced CNN architectures (e.g. AlexNet, VGG-16,etc) may also be considered although they might be too difficult to adequately tune for our problem. This experimentation will be detailed in the next report for Milestone 2.

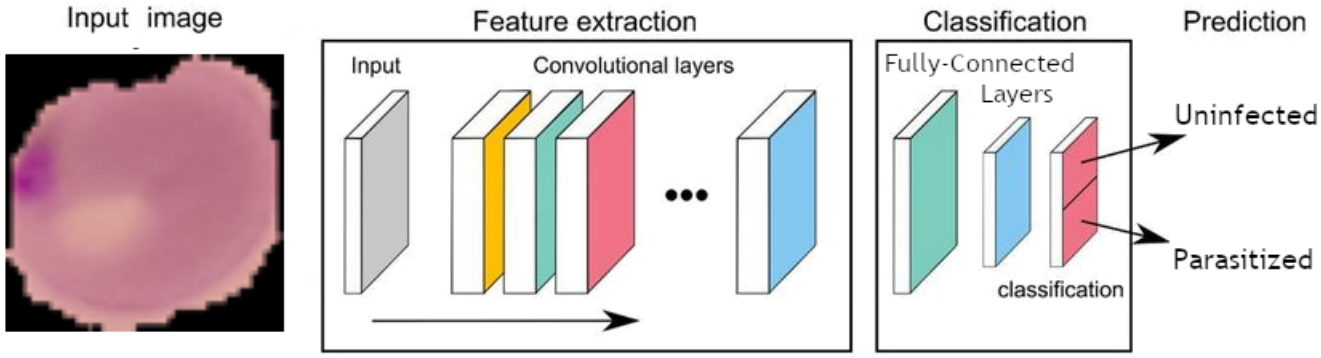

Figure 4: Sketch of the CNN model for malaria detection. Adapted from Huttunen et al. 2019.

3.3 Measures of Success

The detection of malaria from red cells is a binary classification problem, and as such usual score metrics are accuracy, recall, precision, f1-score as well ROC curve or AUC. However, a single of these metrics will be used for hyperparameter search. We will choose recall as our score metric, in order to minimize

false negatives (i.e. parasitized cells predicted as uninfected). This is because we consider the error cost of misdiagnosing a diseased patient to be much greater than the error cost of misdiagnosing a healthy patient. Finally, recall that in practical settings not just one but many red cells (>100) from a patient's blood smear are analyzed, followed by other medical diagnosis. Parasitized red cells are counted which allows to diagnose malaria (above a given treshold) as well as to determine the intensity of the disease. Such a framework is however out of the scope of this project because our dataset only consists only of individual red cells without patient information.

References

- Official CDC DPDx website: https://www.cdc.gov/dpdx/malaria/index.html
- Official NIH Website (for dataset): https://ceb.nlm.nih.gov/repositories/malaria-datasets/
- Huttunen et al., 2019: Investigating and Assessing the Dermoepidermal Junction with Multiphoton Microscopy and Deep Learning. Biomedical Optics Express, DOI:10.1364/BOE.11.000186
- Poostchi et al., 2018: Image analysis and machine learning for detecting malaria, Translational Research, 194, 36-55. DOI: 10.1016/j.trsl.2017.12.004

Appendix: Jupyter Notebook

The next pages are the jupyter notebook used for this report.