Computer Vision for Malaria Detection - Final Submission

Sulian Thual, MIT ADSP November 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 third report (Final Submission), we focus on providing an executive summary, a problem and solution summary, as well as recommendations for implementation.

1 Executive Summary

The goal of this project is to develop a computer vision model for malaria diagnostic from red cell images. Malaria is a disease caused by parasites transmitted through mosquito bites and that infect the red blood cells, causing more than 200 million cases and around 400,000 deaths yearly mostly among children under 5 in Africa. Hundreds of millions of blood films are examined every year for malaria, but manual diagnosis is labor intensive and requires skilled analysts. There is a need for automated diagnosis as proposed here that has the potential of providing more reliable and standardized interpretation while also reducing workload and costs.

Our practical protocol for malaria diagnosis is summarized in Figure 1. A thin blood smear is collected from the patient, from which individual red cells images are extracted. Each individual image is input into the computer vision model in order to predict if the red cell is parasitized by malaria or not. Finally, parasitized cell count allows to diagnose the patient for the presence and intensity of the malaria.

The main work made here has been to build the computer vision model and to envision its use in a mock protocol. We have chosen to use a Convolutional Neural Network (CNN) as our computer vision model. We have used a dataset of around 25000 microscopic red cell images to train the model, augmented 8 folds using image rotations and flips. The computer vision model is very good at predicting if red cells are parasitized: in fact, it reaches 96% accuracy on a separate training dataset of 2600 images without overfitting. We have implemented a statistical analysis (Bernoulli and Binomial distribution) to estimate the accuracy of the practical protocol: while it fails to diagnose patients with early infections ($<$5% parasitized cells), it is good to diagnose patients with more advanced levels of malaria.

Our protocol could be implemented as an automated malaria diagnosis, with potentially potentially enormous medical and economic benefits if deployed at scale but requiring a very careful verification of the protocol to limit risks. One of the key priority for future work is to address cell count errors, either using calibration or trying to increase the CNN accuracy. We found that roughly the same subset of images is systematically misclassified, and suggests to concentrate on its characteristics to develop more suitable feature engineering or data augmentation methods.

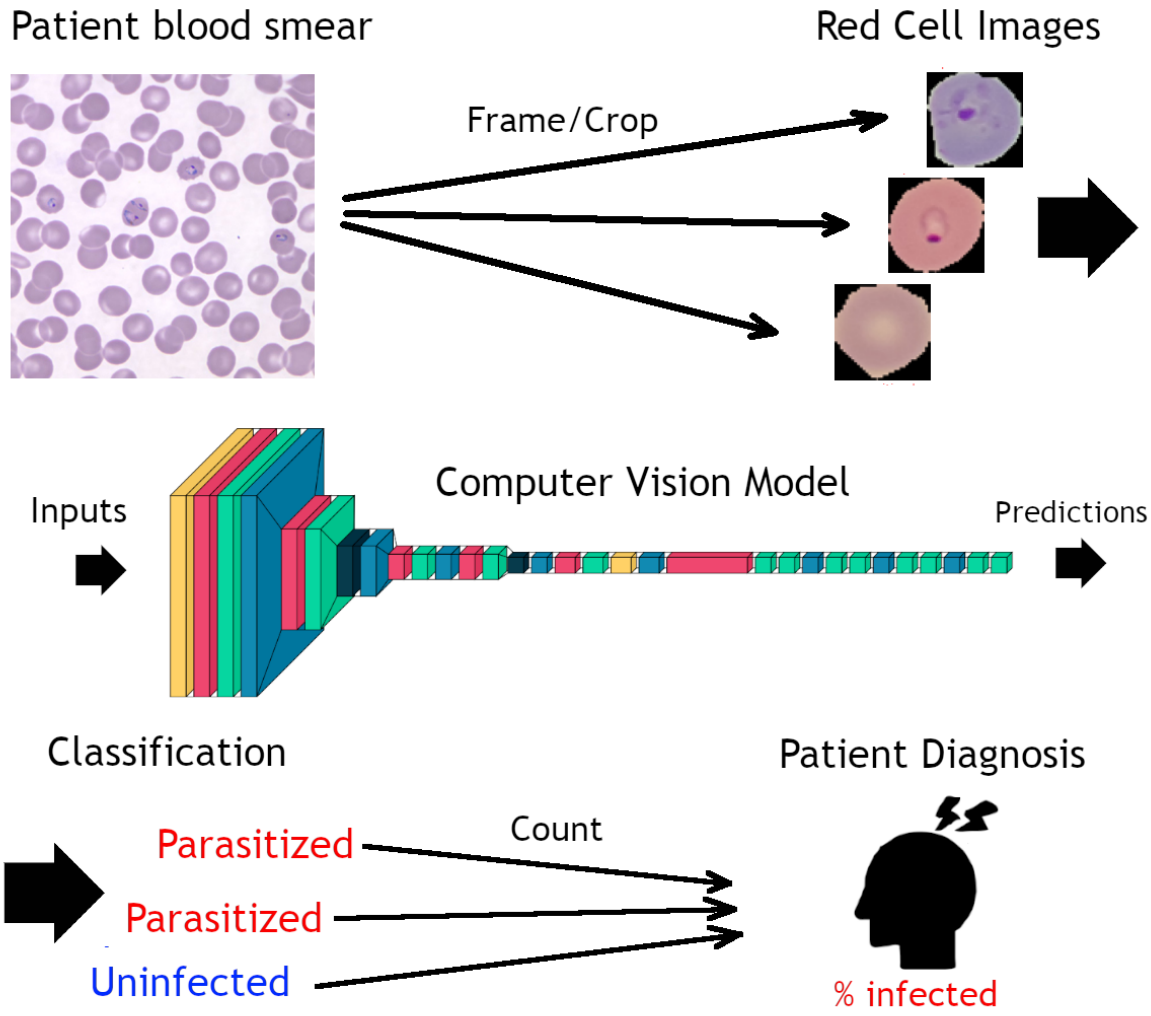

Figure 1: Practical protocol for the automated diagnosis of malaria intensity.

2 Problem and Solution Summary

2.1 Problem Formulation

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 (ranging 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 are perfectly balanced with a 50% split between uninfected and parasitized.

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. Parasitized cells in contrast contain one or several malaria parasites. The parasite can take various shapes (e.g. rings,

elongated, crescents, clefts) which depends on its type 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 look ellipsoidal).

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.

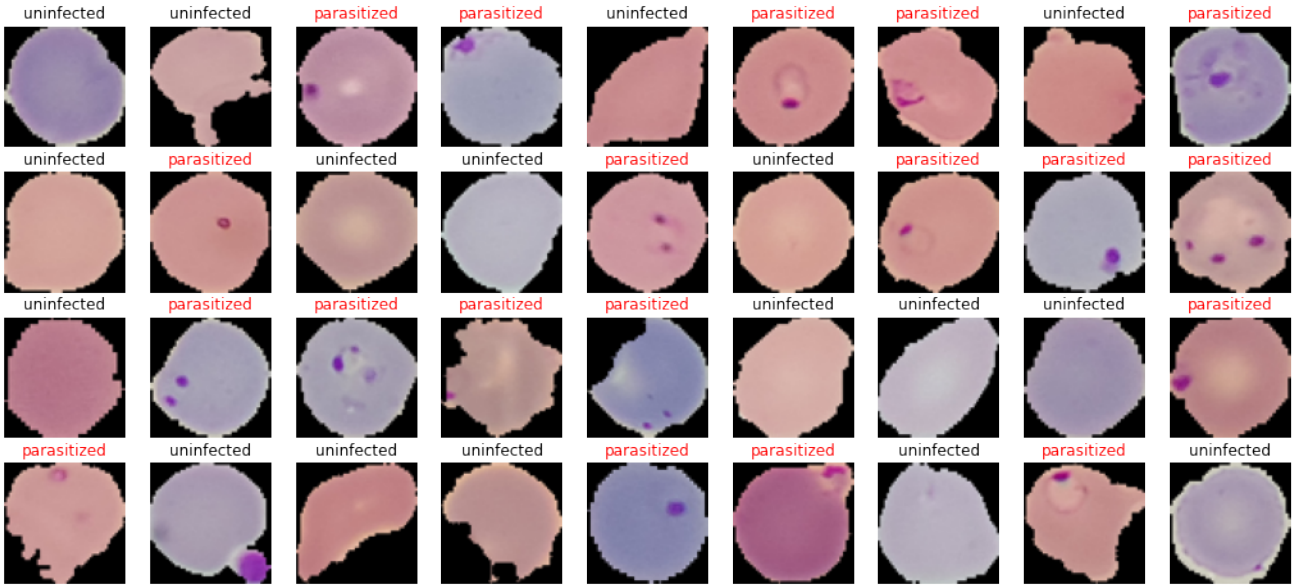

Figure 2: Examples of red cell images from the database (uninfected or parasitized).

2.2 Solution Design

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 artificial neural network most commonly applied to analyze images, and differ from regular Artificial Neural Networks by their use of convolution filters. A major advantage of CNNs is that they use relatively little preprocessing compared to other image classification algorithms, meaning that they optimize the filters (or kernels) through automated learning.

Figure 3 shows the architecture of the CNN model used in this project. Its architecture consists of 5 convolutional layers combined with pooling layers for feature extraction, followed by 4 fully-connected layers (and a final sigmoid activation function) for binary classification. This is a deep CNN with relatively few trainable parameters (around 400 000 weights), which is advantageous because of a relatively short training time (a few minutes).

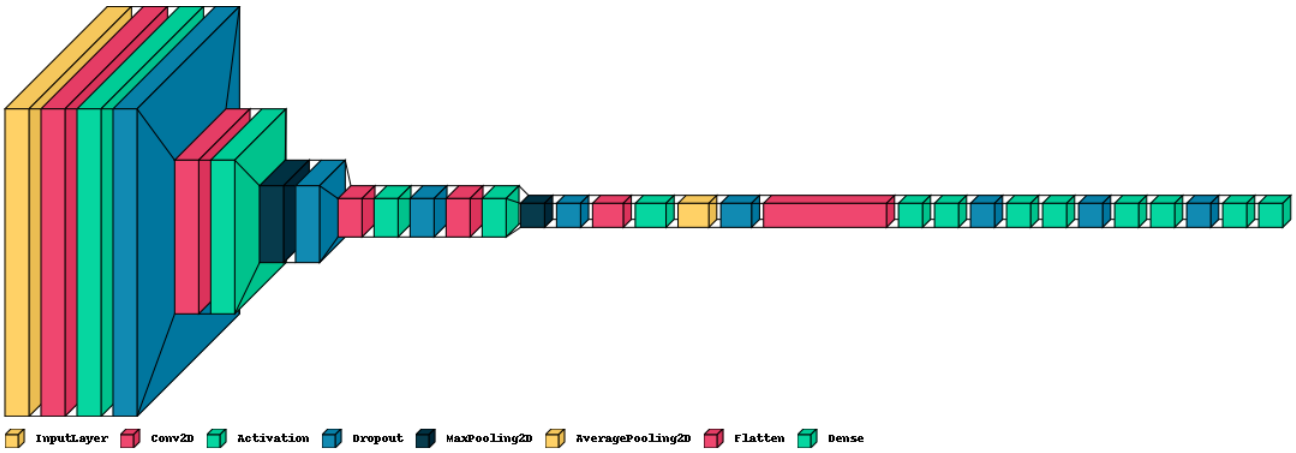

Figure 3: Architecture of the Convolutional Neural Network (using visualkeras).

The CNN model weights were learned using the train dataset. For best performances, we increased the size of the training dataset using data augmentation. For this, we applied transformations as summarized in Figure 4, increasing the training dataset 8 fold by combining 4 rotations x 2 flipped states over any given image. These transformations have the advantage of being "natural", i.e. they could occur randomly in the experimental setup when framing/cropping.

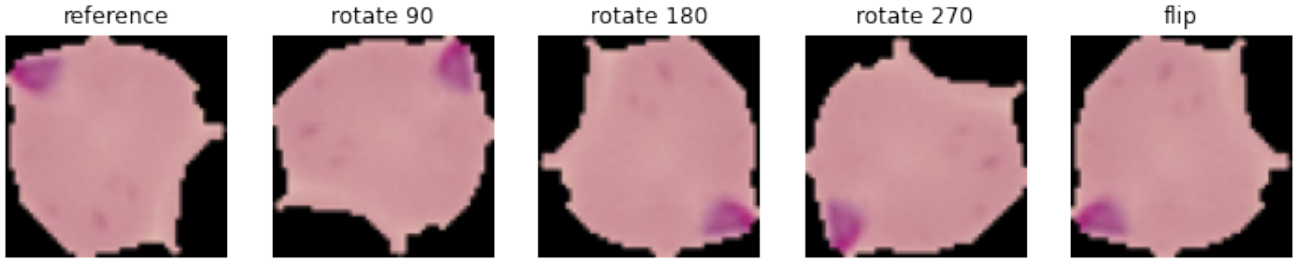

Figure 4: Examples of transformations of a red cell image for data augmentation.

The CNN model performances were assessed on the test dataset, as summarized in Figure 5. The model has good predictive power with accuracy around 96%. Similar values are found for other metrics (precision, recall, f1-score), although these metrics aren't as important for the project goals. We have also compared performances on the train and test dataset to verify that the model is not overfitted (not shown).

Many of the CNN's characteristics have been determined through experimentation not detailed here. We have tested several types of architectures (deep, shallow, pretrained...), data augmentation (using rotations, flips, zooms and shifts) as well as feature engineering methods (e.g. conversion RGB to HSV, editing of image borders, etc). Most models and methods have reached a similar accuracy around 95%, and as a result the best solution was chosen based on its simplicity and ease of implementation (cf Ockham's razor).

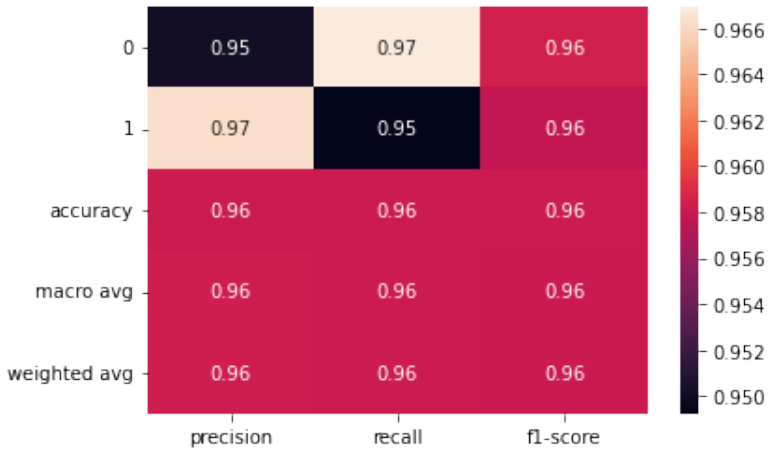

Figure 5: Classification report for model performances on the test dataset (parasitized, uninfected classes are referred to as 1,0 respectively).

3 Recommendations for Implementation

3.1 Takeaways

In this project we have developed a computer vision model (Convolutional Neural Network or CNN) for malaria diagnostic from red cell images. One of the key actionable is the high accuracy (96%) of the model as well as its ease of implementation (e.g. short computation time, etc). Our protocol could be implemented as an automated malaria diagnosis with the expected benefit of providing more reliable and standardized interpretation while also reducing workload. This would have potentially enormous medical and economic benefits if deployed at scale but requires a very careful verification to limit risks.

The key challenges that need to be addressed are as follows. First, the 96% accuracy of the CNN limits the usage of the protocol. In fact, the protocol might struggle to diagnose patients with early infection ($<$5% parasitized cells). One of the key priority for future work is to address this issue. For example, we could calibrate the cell count (changing thresholds for detection) or also try to increase the CNN accuracy. We found that roughly the same subset of images is systematically misclassified that limits the CNN accuracy. It would be interesting to concentrate on the characteristics of such a subset to develop more suitable feature engineering or data augmentation methods. These issues are detailed below.

3.2 Cell Count Error

In our practical protocol for malaria diagnosis (cf Figure 1), a blood smear is taken from a patient containing several (hundreds to thousands) red cells. Each cell is classified by the computer vision model (CNN) then a count of parasitized cells allows to diagnose the malaria. Given that the CNN has only 96% accuracy, this raises the question: what is the error in the count of parasitized cells?

We have implemented a statistical analysis to determine this as summarized in Figure 6. Assuming that each CNN classification is an independent test with Bernoulli distribution, we have shown that the count of parasitized cells follows a Binomial distribution (see details in the code notebook). This formalism allows to compare our protocol performances with the ones of a perfect model with no errors.

As shown in Figure 6, the count of parasitized cells is prone to errors that depend on the computer vision model accuracy (e), the number of samples taken (N) and the patient's level of infection (r). Around 100-1000 samples should be taken to ensure statistical confidence (regardless of the model accuracy). Our protocol with 96% accuracy ($e = 0.04$) fails to diagnose patients with early infection (1% parasitized cells) but is good to diagnose patients with more advanced levels of malaria above 10%. To improve the prediction, we could either calibrate the cell count (changing thresholds for detection) or increase the CNN accuracy.

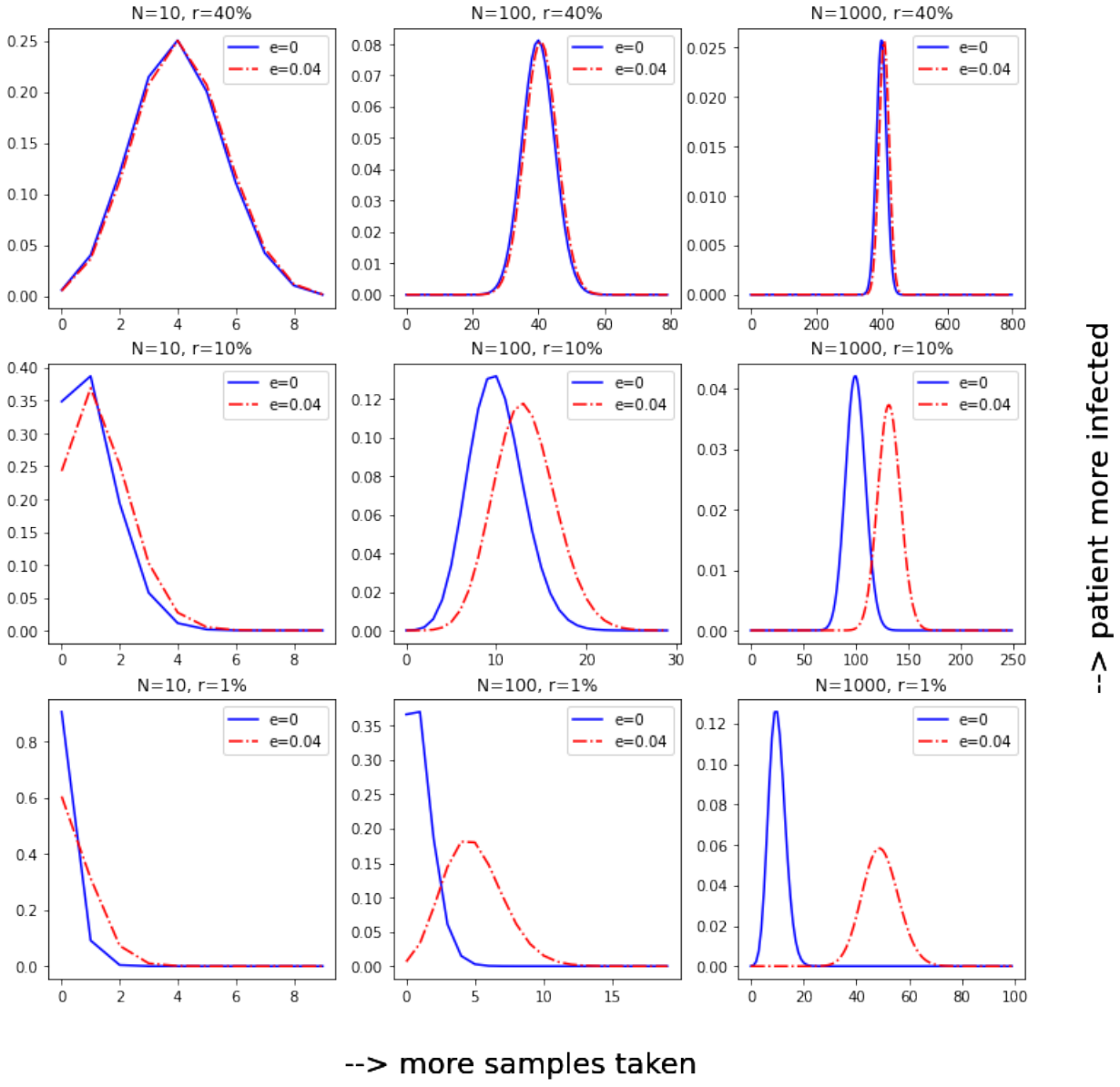

Figure 6: Statistical analysis of the diagnosis protocol. This compares the distribution (probability mass function) of cell counts between a diagnosis that uses the CNN model with 96% accuracy (red) or an (imaginary) perfect model with 100% accuracy (blue). This is repeated for increasing values of the number of sample taken N (left to right) and for increasing values of malaria infection r (top to bottom).

3.3 Misclassified Images

Our computer vision model reaches 95% accuracy, which is also true of most models and methods we have tested. In fact, we have found that roughly the same subset of images (from the test database) is systematically misclassified. How to improve our method to deal with these misclassified images?

Figure 7 shows the subset of misclassified images. Several of these images tend to have a deformed envelope/border with very unique and complex shape. The shape of the envelope is an artifact of the experimental method (framing and cropping) but is irrelevant for malaria detection (in reality, all red cells should be ellipsoidal). Another source of misclassification is having very small parasites, parasites that are located at the very edge of the red cell envelope, or simply very unique cases that even a human observer cannot classify well. It could be interesting to concentrate on these characteristics to develop more suitable feature engineering or data augmentation methods.

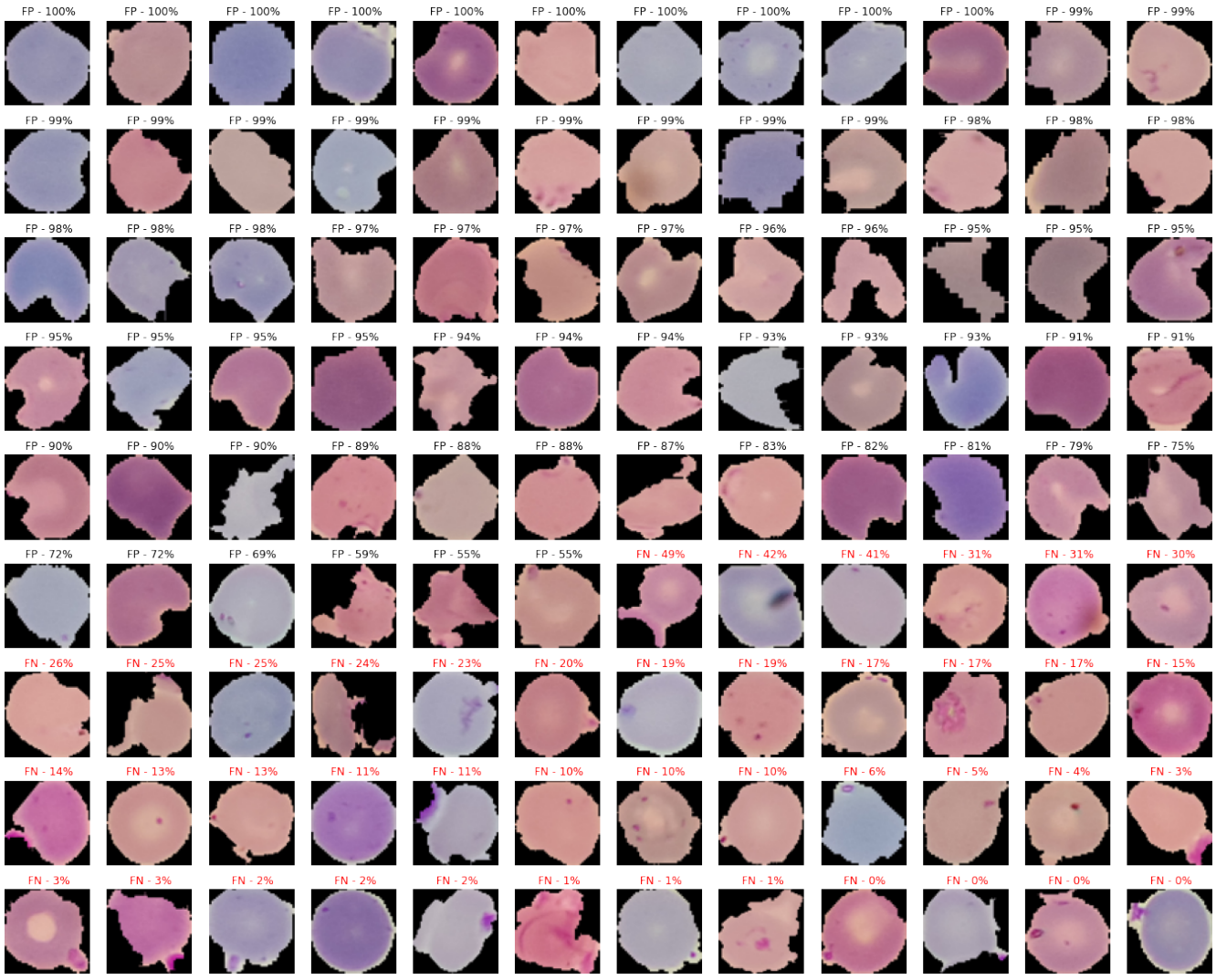

Figure 7: Subset of images from the test dataset misclassified by the computer vision model, with indication of the error type (False Positive or False Negative) as well as the predicted probability of being parasitized (%). Images with p>50% (black) are FP (predicted as parasitized but uninfected), while images with p<50% (red) are FN (predicted as uninfected but parasitized).

Acknowledgements

Special Thanks to Lindsay Morton (Senior Molecular Epidemiologist at DHA) for great discussions about the malaria life cycle and existing protocols for diagnosis.

References

- Official NIH Website (for dataset): https://ceb.nlm.nih.gov/repositories/malaria-datasets/
- Official CDC DPDx website: https://www.cdc.gov/dpdx/malaria/index.html
- VisualKeras library: https://github.com/paulgavrikov/visualkeras

Appendix: Jupyter Notebook

A jupyter notebook is attached to this report, that was used to make all computations and contains implementation details.