A Machine Learning Approach to Detecting Malaria Infection in Blood Cells

Abbie Maemoto, Angela Mao & Grace Soontornviwath

Introduction

Malaria is a life-threatening infection caused by a parasite, which is spread to humans through infected mosquitoes [1]. The gold standard diagnostic tool for malaria is microscopic examination, which involves spreading a patient's blood specimen as a thick or thin blood smear and staining the blood film [2]. The stained blood film is manually inspected through a microscope by malaria microscopists to identify the species and stages of malaria parasites in the blood smear as well as the density of parasites [3]. The manual inspection done by microscopists is tedious and time-consuming, requires experience and extensive training, and is prone to diagnostic errors due to the difficulty of the procedure [4]. There can also be errors caused by the quality of the microscope and staining reagents [4].

This necessitates a replacement of manual inspection by microscopists for a robust cell classification and cell counting machine learning model to reduce human error and the time, resources, and expertise needed, which is the focus of this project. The scope of the project is narrowed down to determining if there are any infected blood cells in the blood smears and does not include identifying the species, stages, and amount of malaria parasites due to the lack of information provided by the image dataset used. The image dataset used to train and test the model is from the "P. vivax (malaria) infected human blood smears" dataset provided by the Broad Bioimage Benchmark Collection [5].

Background

According to the World Health Organization (WHO), there were 249 million reported cases of malaria and an estimated number of 608,000 malaria deaths in the world in 2022 [1]. A disproportionate amount of global malaria cases and deaths falls in the WHO African Region with four African countries (Nigeria, the Democratic Republic of the Congo, Uganda, and Mozambique) accounting for over half of all malaria deaths worldwide [1]. In 2022, the WHO African Region experienced about 94% of all malaria cases and 95% of deaths, making malaria a large concern in this region [1].

In general, early diagnosis of malaria reduces disease, prevents deaths, and contributes to reducing transmission [1]. Earlier diagnosis also leads to patients receiving malaria treatments more quickly, which are determined by the malaria species and density of parasites that are found during diagnosis [4]. However, in resource-poor settings like those in the WHO African Region, the proper healthcare resources and personnel may not be accessible or available to perform the necessary amount of diagnostic testing with adequate accuracy. Therefore, a robust diagnostic machine learning model would reduce the amount of experts needed to perform diagnostic testing and require less time and resources, so more patients can be diagnosed earlier and receive the necessary treatment.

The two most common diagnostic tools used for malaria are microscopic examination and a Rapid Diagnostic Test (RDT). Microscopic examination, as described above, is the gold standard, but requires expert microscopists to manually inspect stained blood smears of patients, which is time-consuming and requires many resources [4]. RDT detects specific malaria antigens

in a person's blood on a test card after 15 minutes [2]. The advantage of the RDT is that it is quick to perform; however, the disadvantages of the RDT are that it may not be able to detect an infection if the blood has lower numbers of malaria parasites, the test is not sensitive to the two less common species of malaria parasites, and a negative or positive result must be verified by a microscopic examination, making the RDT less effective than microscopic examination [2]. This affirms the importance of a quick and robust diagnostic machine learning model that would increase the accuracy and decrease the time of diagnosing malaria.

There are new diagnostic tools currently being created that are based on digital imaging analysis by deep learning and artificial intelligence methods, which include using smartphone applications [4]. However, the implementation of these diagnostic tools in resource-poor settings presents many challenges, such as determining how to implement the technology in regional hospitals or small healthcare centers [4]. With this project we hope to take the first step in creating a technology that could be easily implemented by creating an accessible machine learning model that can be performed on a laptop.

Methods

Dataset Overview

The malaria dataset from Broad Bioimage Benchmark Collection contains a train.json and test.json, each containing the labels for each blood sample image. There are a total of 1208 images in the training set and 120 images in the test set. Each image represents an individual row in the file; within an image, each cell is labeled with the following fields: “minimum”, “maximum”, and “category”. The “minimum” field contains the coordinate of the location of the

upper left corner of the cell. The “maximum” field contains the coordinate of the location of the bottom right corner of the cell. The “category” field contains the type of blood cell, including “red blood cell”, “leukocyte”, “gametocyte”, “ring”, “trophozoite”, and “schizont”.

Malaria Data Preprocessing

Based on the “minimum” and “maximum” coordinates, we first calculated the area of the cell. We would like to acknowledge that given these 2 coordinates, only the rectangular area could be calculated, overestimating the exact area of the cell. Additionally, the paper stated that cells labeled “red blood cell” or “leukocyte” were not infected with malaria; cells labeled “gametocyte”, “ring”, “trophozoite”, or “schizont” were identified to be infected with malaria. We therefore added another field in the dataset called “diagnosis” where cells labeled “red blood cell” or “leukocyte” had “diagnosis = 0”; otherwise, the cells were labeled “diagnosis = 1”.

Image Analysis for Preprocessing

Since the initial data only came labeled with coordinates and whether or not a cell was infected, we decided to self-label some images to create an improved dataset with additional proxies that could potentially enhance model performance. To do so, we took 9 images from the dataset, and self-labeled a random sample of 20 cells from each to create a new dataset of 180 cells. To analyze the images, we used ImageJ. We first opened each individual image, converted the images to 8-bit grayscale, enhanced the contrast, and added a threshold to isolate the cells from the background. We also used Process → Noise → Despeckle to remove any noise/small particles. Then, we did Analyze → Analyze Particles, and checked Display Results, Clear Results, and Show Outlines, which outputted the areas for each segmented cell. This method of

calculating area is more accurate than the area calculation for the Broad Bioimage Benchmark Collection malaria dataset because ImageJ calculates exact area versus the rectangular area calculated from the “minimum” and “maximum” coordinates.

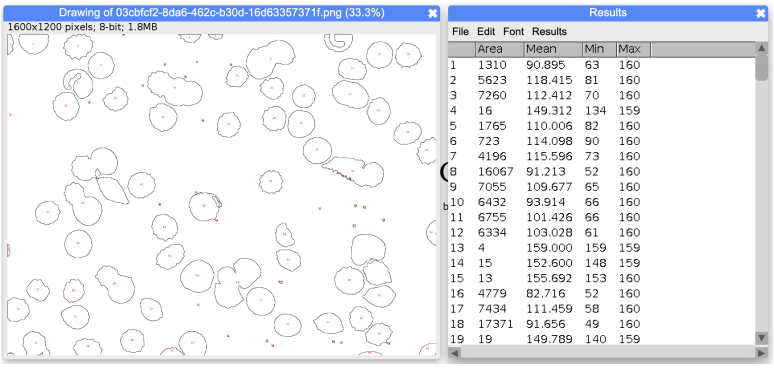

Figure 1. Cell areas generated from ImageJ

We also looked at the unaltered photos and visually classified the cells by shape (smooth circle, bumpy circle, spiky circle) and color (blue or purple).

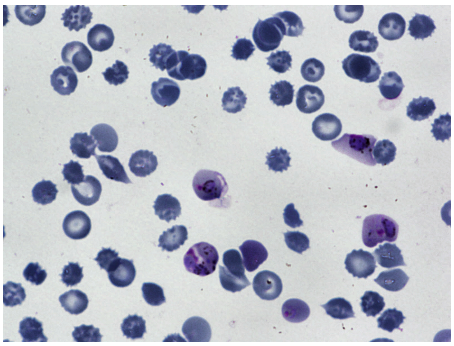

Figure 2. Raw images of the RBCs, purple indicating infected cells

To build our more detailed dataset, we created a new CSV that contained the area, shape, color, and infection status (infected or not infected). We planned to use this new data to train another model, since we believed the area measurement done by ImageJ would be more accurate than the one done through coordinate estimation.

Code

The goal of our models are to classify whether or not a cell was infected with malaria based on various features. In Model A, we trained on the train.json from the Broad Bioimage Benchmark Collection malaria dataset, and used the rectangular area as the sole feature for prediction. In Model B, we trained on the hand-labeled dataset and used exact area as the sole feature for prediction. In Model C, we trained on the hand-labeled dataset using exact area and shape as the two features for prediction. Because we are solving a classification problem, we selected a logistic regression machine learning model for training and testing. We utilized the sklearn Logistic Regression model and metrics libraries to implement the software solution. The CoLab for Model A can be viewed here. The CoLab for Model B can be viewed here. The CoLab for Model C can be viewed here. To comparatively evaluate whether or not blood cell area alone is an effective proxy for detecting malaria, we compared the performance metrics of accuracy, precision, recall, f1 score, and ROC AUC between the three models. These metrics are measures of the model's prediction rates of true positives, false positives, true negatives, and false negatives.

Results

Model A:

ROC AUC	Accuracy	Precision	Recall	F1-Score
0.80	65.70%	10.75%	76.62%	18.86%

Model B:

ROC AUC	Accuracy	Precision	Recall	F1-Score

0.94	91.67%	66.67%	80.00%	72.73%

Model C:

ROC AUC	Accuracy	Precision	Recall	F1-Score
1.00	94.44%	71.43%	100.00%	83.33%

Analysis

As seen in the results above, the metrics of ROC AUC, accuracy, precision, recall, and f1 score improved when using exact area (Model B) versus rectangular area (Model A) as the predictive feature. These metrics further improved when using both exact area and shape (Model C) as predictive features versus exact area alone (Model B).

For Model A, the metrics of ROC AUC and recall performed best while the metrics of precision and f1 score performed worst comparatively. This indicates that there was a low ratio of correctly predicted positive instances (malaria positive) to the total predicted positive instances, highlighting a relatively high false positive rate by the model. On the other hand, the model had a low false negative rate, which is an important signal since false negatives in this case mean that the model predicted that the cell is not infected with malaria when the cell is indeed infected in reality. For Model B, we observe that the metrics of recall, and hence f1 score, are significantly improved which informs us that the accuracy of area calculation (rectangular versus exact) plays a critical role in the logistic regression model's ability to correctly predict positive cases. For Model C, we observe that the introduction of shape as a feature, in addition to exact area,

improves all 5 metrics as compared to Model B. This allows us to arrive at the conclusion that blood cell area alone is not the best proxy for detecting infection; rather, blood cell area and shape significantly improves the predictive power of the model. Future studies will e

Lastly, we would like to acknowledge limitations of our study. The Broad Bioimage Benchmark Collection dataset had 1208 images, each with dozens of cells; thus, Model A was trained on over 80,000 cells. Comparatively, Model B and C were only trained on 180 cells due to the time intensive nature of hand labeling. Additionally, when hand-labeling the blood cells, we considered using color of the cell (light blue, dark blue, purple) as a feature. However, after inspection of the images, we realized that different staining methods were used for different images, impacting the color of the cell. Without a consistent color baseline for labeling, we decided to exclude color as a feature for prediction in the model.

In future studies, we would like to use ImageJ/Fiji to automate the labeling process such that we can automate the labeling process and train models B and C on a more comparable dataset as model A's. We would also like to investigate how other features (cell texture, density, etc) impact the predictive power of the model, and how other machine learning models (Random Forest Classifiers, neural networks, etc) perform on the same dataset.

Contribution

All three members contributed equally to the research and conception of the idea. Angela and Grace worked on the image analysis portion, while Abbie worked on the development of the models.

References

1. *Fact sheet about malaria*. (n.d.). Retrieved December 4, 2024, from https://www.who.int/news-room/fact-sheets/detail/malaria
2. CDC. (2024, May 10). *Malaria Diagnostic Tests*. Malaria. https://www.cdc.gov/malaria/hcp/diagnosis-testing/malaria-diagnostic-tests.html
3. *Microscopy examination of thick and thin blood films for identification of malaria parasites*. (n.d.). Retrieved December 4, 2024, from https://www.who.int/publications/i/item/HTM-GMP-MM-SOP-08
4. Maturana, C. R., de Oliveira, A. D., Nadal, S., Bilalli, B., Serrat, F. Z., Soley, M. E., Igual, E. S., Bosch, M., Lluch, A. V., Abelló, A., López-Codina, D., Suñé, T. P., Clols, E. S., & Joseph-Munné, J. (2022). Advances and challenges in automated malaria diagnosis using digital microscopy imaging with artificial intelligence tools: A review. *Frontiers in Microbiology*, *13*, 1006659. https://doi.org/10.3389/fmicb.2022.1006659
5. *Broad Bioimage Benchmark Collection*. (n.d.). Retrieved December 4, 2024, from https://bbbc.broadinstitute.org/BBBC041