Capstone Project

Malaria Detection - By Nishad Khudabux

Context

Malaria is a contagious disease caused by parasites that are transmitted through the bites of infected mosquitoes. The parasites enter the blood and begin damaging red blood cells (RBCs) that carry oxygen, which can result in respiratory distress and other complications. Late treatment can cause complications and could even be fatal, but as the symptoms can mimic the flu and stay dormant for a year or more, **early detection is key**.

Malaria is known as a **disease of poverty**. Despite treatment being **highly effective** and **relatively inexpensive**, almost 50% of the world's population continues to be at risk from malaria. There were more than 229 million malaria cases and 400,000 malaria-related deaths reported over the world in 2019. This challenge primarily stems from a **bottleneck that makes early detection difficult** in poorer parts of the world:

- Collecting blood samples requires the use of a large heavy centrifuge and microscope, often in remote areas that do not have access to electricity.
- **Diagnosis of those samples requires inspection by experienced professionals (which also needs to be factored into the cost). Given the magnitude of the problem, this is a repetitive and time-consuming process.**

While the first problem has largely been solved through inspired innovation (paperfuge and foldscope), the second continues to be an issue. However, an automated model can be an innovative solution. Not only providing **faster** and **cheaper** diagnoses but also with **greater accuracy** than found with professionals.

In the words of Manu Prakash from Stanford's frugal labs, and the inventor of foldscope and paperfuge:

"Every 0 that you add to the cost of scientific equipment, probably thousands and millions of people are cut off"

An automated system using machine learning to perform malaria detection has the potential to have a **tremendous positive impact** on the global population.

Objectives

Build an efficient computer vision model to detect malaria. The model should identify whether the image of a red blood cell is that of one infected with malaria or not, and classify the same as parasitized or uninfected, respectively.

Key Questions

- Load the data. Are there any corrupted files?
- Is the division of training to test data appropriately?
- Is the division of positive to negative cases well balanced?
- Note observations on the images
    - What features could the model potentially recognize to classify parasitized and uninfected cells?
    - What features could potentially cause errors in the model?
- Test different types of image transformations, which would increase the model's accuracy?
- What are some potential model design solutions, and how do we measure if they are overall successful?

Problem Formulation

Build a computer vision model that can work in a larger automated system that can provide faster, cheaper, and more accurate malaria diagnoses. Tackling one of the world's most deadly diseases in poorer parts of the world that require innovative solutions.

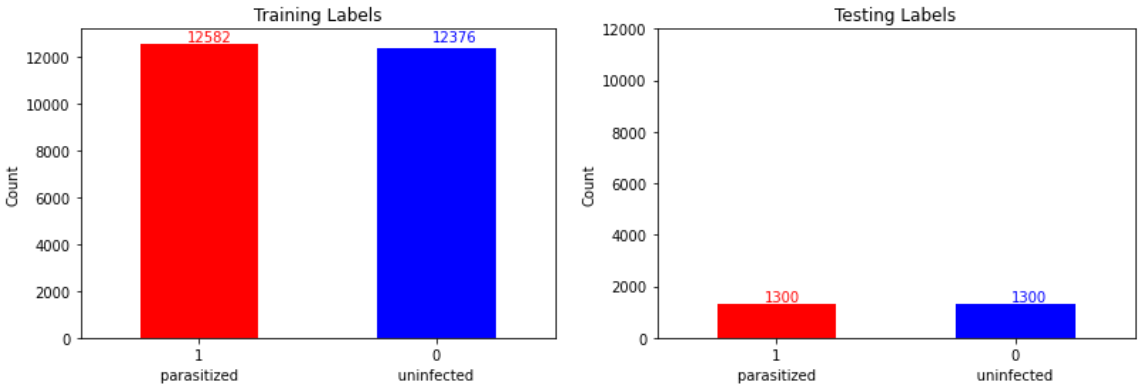
Data Exploration

Data Description

There is a total of 24,958 training and 2,600 testing images (coloured) that we have taken from microscopic images. These images are of the following categories:

Parasitized: The parasitized cells contain the Plasmodium parasite which causes malaria

Uninfected: The uninfected cells are free of the Plasmodium parasites



As shown above, the training testing split is appropriate for building neural networks (about 10 to 1), and the number of images in both classes is well-balanced (about equal). This should lead to a balanced, properly trained/tested model. All images were properly loaded.

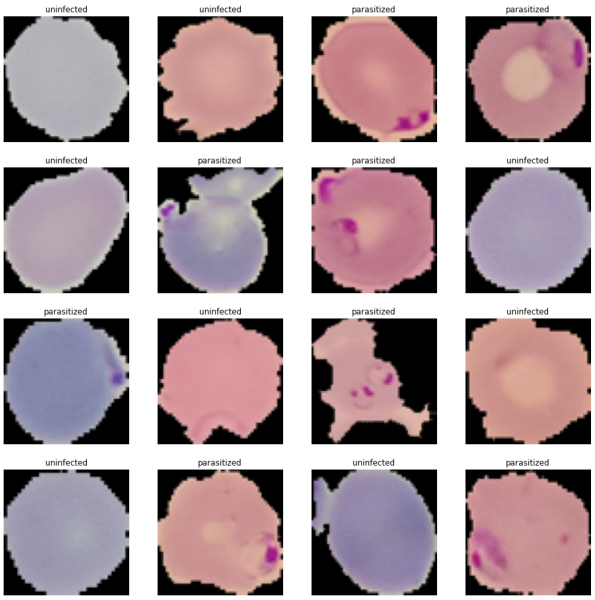
Data Exploration

Our data contains images of isolated red blood cells.

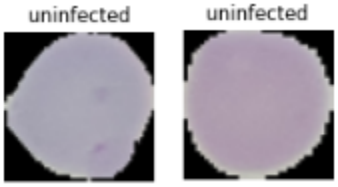
Some have not been cropped perfectly, but as long as it has not cropped out the infected area this should not be an issue. Also if this model is part of a larger system that weighs multiple cells in a blood sample, that would mitigate any cropping errors.

They have both blue and pink backgrounds (looks like slightly more pink cells in total). Unsure as to why some have a blue background as most microscope blood smears show pink cells only?

Infected cells are **identifiable by the presents of parasites** in the red blood cells. They can be recognized by:



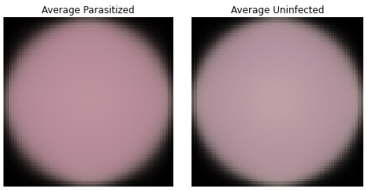
- **Color** - Taking on a distinct **purple** hue. This has the potential to cause false positives as some of the uninfected pink cells contain areas of darker colour similar to the parasite colour.
- **Shape** - The grouping of the parasites tends to form distinct shapes inside of the cell (ring, band, rosette, dots, feathering). These shapes are potentially recognizable in Convolutional Neural Networks using a combination of filters (edge detections) and pooling.
    - The difference in shape is key in diagnosing the different types of malaria (*Plasmodium falciparum, Plasmodium malariae, Plasmodium ovale, and Plasmodium vivax*). While we are simply looking for positive and negative cases here, there is potential to train future models to classify the type of malaria present.



Upon further inspection, a small number of uninfected cells showed minimal signs of what looked like purple specs. This might indicate that there is potential for **noise** to give us false positives.
To the left are 2 examples found in the 6x6 subplot

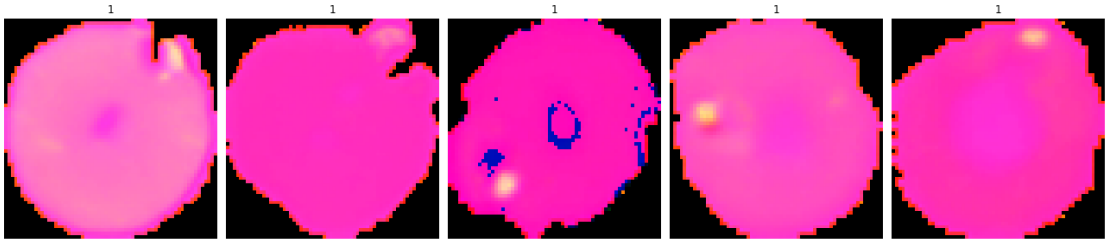
Image Transformation

We attempted three forms of image transformation (Image Averaging, HSV, and Gaussian Blurring) to hypothesize if any had the potential to positively affect the accuracy of our model.



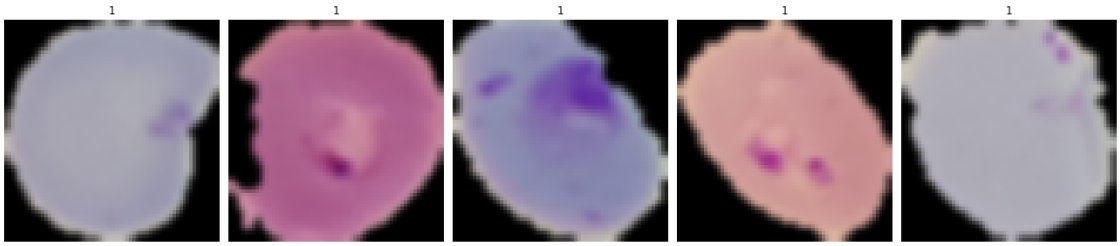
Averaging of Parasitized and Uninfected Cells

There is very little to distinguish the average parasitized image from the average uninfected image, making it a **poor tool to draw any insight**. This is not surprising as the majority of the parasitized and uninfected cells are similar in colour and shape. Only the presence of the parasites (relatively small area of the cell) is different, and that is lost in the blending of colours.



Conversion from RGB to HSV

The HSV images are much more pronounced in their colour differences. This would likely prove **easier for a neural network to distinguish**, resulting in greater accuracy. It would be prudent to test if converting our images to HSV improves our model.



Processing images using Gaussian Blurring

The primary reason for using Gaussian blurring is to **reduce the noise** in images. As we stated above, there is **potential for noise to cause false positives** in uninfected cells. We can still clearly identify the shape and colour of parasitic areas, and therefore the reduction in image sharpness may not result in a decrease in accuracy. While the noise reduction can potentially result in a decrease in false positives. If our models show signs of higher than expected false positives we should consider processing our images using Gaussian Blurring.

Given that the parasitic areas are generally darker than the colours around it, it would be interesting to see if a model would do well on **grayscale images** (especially if we could increase contrast to increase that difference). It would also decrease the size of the data and potentially save computational resources.

Proposed Approach

Potential techniques to explore:

As with most image recognition, adding **more data** will generally increase validation accuracy.

- By finding other databases of labelled malaria images and joining with our current database.
- Through **data augmentation** to increase the diversity of our database and potentially create an even stronger model.

As well as implement some form of **transfer learning** that generally has positive results on overall accuracy.

It should be noted however that all this comes at the cost of increased complexity and computation. It is generally a good rule not to over-engineer a complex model when a simpler model can achieve the same thing. It is therefore better to fine-tune these parameters to find the optimal balance of performance.

Overall solution design:

As we observed above the infected cells can be identified primarily through **colour** and **shape**. If colour is the primary means of classification, then it is possible that an ANN model would work. However, to take advantage of both a **CNN model** would likely do much better.

From there it would be a matter of optimizing the **hyperparameters** and trying **different architectures** to see which is best. Before we even begin, we can likely anticipate that:

- Some **pooling** layers will be beneficial to recognize the parasitic shapes
- We should use the **Relu activation function** as it performs best for computer vision
- We should consider using **recall** as our performance metric, as the cost of a false negative could potentially lead to death. However, if we are trying to be conscious of increased costs from false positives, using **accuracy** as our performance metric is arguably another option.

Measures of success:

Key measures of success will be if we can produce a model that gives high **validation accuracy** and **generalized performance**. That is to say; the model is not overfitted (does as well or better on the validation data as it does on the training data).