# Capstone Project: Detection of Malaria using Convolutional Neural Networks

Report prepared by: Ethan Hughes

### Executive Summary

This project proposes the use of a Convolutional Neural Network (CNN) for the detection of malaria by identifying malaria infected red blood cells. The suggested final model has high accuracy and a low rate of false negatives which could greatly impact early detection rates if implemented successfully. Despite promise, the model has limitations that may affect its practical performance that stem from the image data the model was trained on. It is recommended that stakeholders determine the performance of the model against full blood smear images before investing in widespread implementation. Overall, the proposed model may serve as a valuable pre-trained model that can accelerate the development of a practical model that increases early detection of malaria thus decreasing malarial mortality worldwide.

### Problem Summary

Malaria is a life-threatening vector borne disease that is common in tropical regions around the world. According to the World Health Organization (WHO), there were an estimated 241 million cases of malaria worldwide in 2020; 627,000 of which were fatal (Malaria Fact Sheet, 2022). Currently approximately 50% of the world’s population lives in areas at risk of malaria transmission. Unfortunately, this number is expected to grow as climate change further expands the habitable area of the vector that transmits Malaria (Fernando, 2022).

Malaria is contracted when an infected Anopheles mosquito (the vector) bites a human and passes Plasmodium parasites into the victim’s bloodstream. Once in the blood, the parasites damage red blood cells (RBC) which can lead to respiratory distress, fever, chills, headaches, fatigue, and other complications. Malaria can rapidly become severe, but if it is detected early, it is easy to treat. Antimalarial drugs are relatively cheap and easy to store; however, complicated cases require treatment beyond simple antimalarial drugs. Therefore, detecting infections prior to complications is essential for reducing malarial mortality.

Malaria is detected by analyzing blood cells under a microscope to look for infected cells. Unfortunately, it requires highly trained professionals to identify infected cells under a microscope, and thus, detection can be difficult - especially in underserved areas that might not have access to skilled professionals. One potential solution to improve malaria detection is to use machine learning and artificial intelligence to identify infected cells.

The key objective of this project is to build or identify a convolutional neural network (CNN) that can accurately identify red blood cells infected with malarial parasites. The hope is that this model can then be used to boost early detection rates and reduce malarial mortality worldwide.

### Solution Design

As part of the solution design, five different models were explored: (1) a simple CNN model, (2) a more complex CNN model with additional layers, (3) a CNN model using batch normalization and a LeakyRelu activation, (4) a more complex CNN model with additional layers and data augmentation, (5) transfer learning using a pre-trained VGG16 model. The final proposed model is a deep learning CNN model with 5 convolutional layers, 5 pooling layers, and 3 dense layers which has been optimized with hyperparameter tuning. The structure of the final model can be seen in **Figure 1**. (It is worth noting that a fully convolutional network (FCN) was not tested for two reasons: 1. My understanding is that FCNs perform better with semantic segmentation, and the goal of this project is image classification, 2. The goal of the project is specifically to test a neural network, which is a component missing from a FCN.)

Figure : Final Convolutional Neural Network Model

Table

Description automatically generated

The final model has high accuracy (97.82% for training data and 98.31%for test data). The final model also has a low number of false negatives (11), and consequently, very high recall (99.15%). The final model had the highest accuracy of all the models ran, however all accuracies were similar (**Table 1**). The main differentiator with the final model was the lowest number of false negatives compared to other models. Both the final model and model 2 have low false negatives. Given that false negatives are the costliest outcome, reducing the number of false negatives is a priority which is why the final model was selected.

Table : Performance of different models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Train Accuracy | Test Accuracy | False Negatives | False Positives | True Negatives | True Positives |
| (1) Simple CNN | 97.96% | 97.65% | 29 | 32 | 1268 | 1271 |
| (2) Complex CNN | 98.01% | 98.27% | 11 | 34 | 1266 | 1289 |
| (3) CNN w/ BatchNormalization and LekyReLu | 96.89% | 97.50% | 58 | 7 | 1293 | 1242 |
| (4) CNN w/ data augmentation | 97.83% | 97.92% | 28 | 26 | 1274 | 1272 |
| (5) Pretrained model | 96.08% | 94.73% | 56 | 81 | 1219 | 1244 |
| (6) Final model | 97.82% | 98.31% | 11 | 33 | 1267 | 1289 |

When constructing the final model, the format was based off the Complex CNN model as that model performed better than models 1, 3, 4 or 5 with regard to false negatives. To create the final model, experimentation was done by adding a few additional layers, adding more dense layers, and hyperparameter tuning until the lowest number of false negatives was obtained. Although both the complex CNN model and the final model have essentially the same performance, I suspect the final model will do better in practical implementation (see Recommendations for Implementation section below).

### Analysis and Key Insights

The final model has high accuracy when determining if a single cell is infected with the parasites that cause malaria. Addition of new convolutional layers did not positively impact the performance of the model; likely because the images have few features that need to be detected to determine if the cell is infected. This might not be the case in a practical setting where the model will be expected to look at blood smears instead of images of single red blood cells (see limitations section below).

The final model validation accuracy and test accuracy are close to one another, which is a good indicator that there are not overfitting problems (**Figure 2**). The validation accuracy does fluctuate across Epochs (first performing better than training, then worse, then increasing back to near the training accuracy), but does appear to be converging with the training accuracy. Additionally, while the validation accuracy does appear to have large fluctuations due to the scale of the axis, note that the actual fluctuations are relatively small (approximately 2%).

Figure : Performance of final model as measured by accuracy for training and test data

Chart, line chart

Description automatically generated

In addition to high accuracy, the final model has a low number of false negatives when run against the test data (**Figure** 3).This is important because correctly identifying an infected cell is critical to early detection and can minimize the severity of malaria. The false positive rate is also low. While false negatives are more important, we should keep false positives in mind as anti-malarial drugs are not side-effect free.

Figure : Performance of final model as measured by predicted counts versus actual counts

A picture containing chart

Description automatically generated

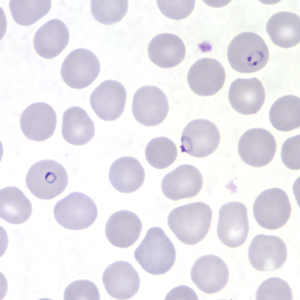
Outside of performance, it is tough to determine the key insights the model is detecting to determine its classification (parasitized vs uninfected) of each cell. This is why neural networks are often considered “black box” models, because the factors driving prediction remain obscured. However, this is not a limitation of the model, because the factors driving prediction are not the important insights we are hoping to learn. Our primary insight is whether malaria was properly detected.

### Limitations and Recommendations for Further Analysis

While the final model has high accuracy in determining if a single RBC is infected with the parasites that cause malaria, there are a few limitations that persist and should be explored, including: practicality of diagnosis, success across the 3 strains of malaria and differential diagnoses, and the benefit cost of implementing this solution instead of preventative measures.

If the goal is to increase the early detection rate of malaria, there are many challenges that may prevent this model from accomplishing that goal. The main challenge is that in a practical setting, a blood smear is the sample collected from a patient used to diagnosis malaria. A blood smear contains many blood cells on a single slide, not just a single cell (**Figure 4**). This model performs well determining if a single cell is infected, but it is unclear how well the model would perform when looking at multiple cells simultaneously. For example, if a single smear has 25 RBCs, and only one is infected, will the model recognize that this patient has malaria? This challenge can likely be addressed in the implementation of a working model, but the performance against blood smears remains unknown.

Figure : Blood Smear, example of clinical presentation used to diagnosis malaria.



Another unexplored challenge with this model is that its performance against individual strains of malaria is unknown, as well as its performance in the presence of comorbidities. There are three strains of Malaria, and the presentation of each is not the same. This model was trained on malarial infected cells, but the type of malarial infection is not specified. Therefore, it may perform better against some strains than others. Additionally, if determining the type of malarial infection is important, it is unclear how well this model performs in differentiating the different strains. Finally, there are other conditions that can affect red blood cells (and other cells that appear on a traditional blood smear). This model may incorrectly diagnose cells with comorbidities as parasitized when the cells might not be infected by malarial parasites.

Finally, there are a variety of methods available to prevent the spread of malaria. A full health impact assessment should be performed comparing the cost-benefit of implementing this solution as opposed to funding preventative measures.

### Recommendations for Implementation

As climate change puts more people in malaria-at-risk climates, a tool that can quickly, and accurately identify malaria at scale will become vital in reducing malaria-related deaths. The proposed model solution shows promise in correct identification of malaria infected red blood cells and has potential applications in testing worldwide. If successfully implemented, this model may help increase early detection efforts and reduce the global malarial mortality.

The proposed model is likely not ready for direct field implementation as its performance against a clinical blood smear remains untested. To successfully implement this model, I recommend treating this proposed model as a pre-trained model and running it against a new set of training and test data that is composed of correctly labeled blood smears. Retraining the model against the images it will likely be using to diagnose malaria is critical to success in the field. Without doing so, we have no understanding of how successfully the model will perform in a real-world clinical setting. The model may perform terribly against blood smear slides, in which case malaria diagnoses may be missed, and lives will be at risk. One reason the final model was selected over model 2 was that the additional convolutional layer and dense layer may help with both identification and interpretation of additional features present in the blood smear slides compared to the single cell images.

One final risk of implementing this solution is that there may be more cost-effective solutions targeting malarial prevention. Running this model might not be feasible for all regions of the world and may be expensive for underserved areas. There are numerous low-cost solutions aimed at preventing malaria, including education campaigns around standing water, bed nets, and proper pest control. It might be more cost effective to invest in these solutions. For this reason, a full health impact assessment should be performed before the model is implemented.

# Bibliography

Centers for Disease Control and Prevention. (2020, October 6). *Image Gallery - Laboratory Identification of Parasites of Public Health Concern*. Retrieved from CDC.gov: https://www.cdc.gov/dpdx/malaria/index.html

Fernando, S. (2022, June 8). *Climate Change and Malaria - A Complex Relationship*. Retrieved from United Nations: https://www.un.org/en/chronicle/article/climate-change-and-malaria-complex-relationship

Organization, W. H. (2022, june 8). *Malaria Fact Sheet*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/malaria

**Special thanks to Dr Adam C. Huseny, MD for consultation about practical limitations**