Capstone Project: Malaria Detection

Final Report

Joon Jung

Executive Summary

This project proposes the use of Convolutional Neural Networks (CNNs) with data augmentation model for the classification of cell images to detect Malaria. The suggested model shows high accuracy and low false negative rate, a specific need that we had for this project. In between two different types of error that model can have, false negative and false positive, we put more emphasis on lower false negative rate than false positive rate for this project in particular, since false negative result in Malaria detection could lead to Malaria patient being left untreated for a long period of time that can be fatal to a patient even if patient gets treatment later on. This model satisfies not only high accuracy and stable validation result, but also low false negative rate compared to various different models that were tested with.

Problem Summary

Malaria is a contagious disease that causes more than 229 million malaria cases and 400,000 malaria-related deaths reported over the world in 2019. Almost 50% of the world’s population is in danger from malaria and children under 5 years of age are the most vulnerable population group affected by malaria as they accounted for 67% of all malaria deaths worldwide. Any contribution of reducing the numbers of malaria-related deaths or streamlining diagnosing process can benefit millions of people and it is a very important problem that we must face and solve.

The pathology of malaria is that it is a contagious disease caused by Plasmodium parasites that are transmitted to humans through the bites of infected female Anopheles mosquitoes. The parasite enters the blood and begin damaging red blood cells (RBCs) that carry oxygen, which can result in respiratory distress and other complications. The lethal parasites can stay alive for more than a year in a person’s body without showing any symptoms, which means that late treatment can cause complications and could even be fatal. This leads to the fact that early, fast, and accurate diagnosis of malaria is very crucial. However, the problem rises where the traditional diagnosis of malaria in the laboratory requires careful inspection by an experienced professional to discriminate between healthy and infected red blood cells. It is a tedious, time-consuming process, and the diagnostic accuracy can be adversely impacted by inter-observer variability since it heavily depends on human expertise.

This is where a data science can provide a solution. An automated system using automated classification techniques using Machine Learning (ML) and Artificial Intelligence (AI) have consistently shown higher accuracy than manual classification by human, and it can drastically help with the early and accurate detection of malaria. It would be beneficial to propose an efficient computer vision model that performs malaria detection using Deep Learning Algorithms, where it can 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.

Solution Design

Several different models and methods were explored as part of the solution design, such as advanced base model, model using Batch Normalization, model with data augmentation, and well-known pre-trained model. The final proposed solution is the model with data augmentation, which showed promising results in high accuracy, high validation accuracy, and low false negative rates. In this particular method, data augmentation such as horizontal flip, zoom in, and rotation has been used as an image augmentation method.

Chart

Description automatically generated

Figure 1: Best overall model’s accuracy result with data augmentation

The base model, illustrated in Appendix 1, also produced a very good and close match to the performance of this model. While it is a solid robust base model to work with, this method showed more stable validation accuracy as the model trains over numbers of epochs, and it showed lower false negative rate which was important for this project in particular. Therefore, I would recommend using model with data augmentation for more suitable and reliable result for our project.

Analysis and Key Insights

Let’s draw more insights into our proposed model with data augmentation. For this particular model, a combination of horizontal flip, zoom in, and rotation was used. This is the visualized images of the augmented images.

Background pattern

Description automatically generated

Figure 2: Augmented images used in proposed model

These images may seem odd in our human eyes, but it is different how our models see them. Other types of image augmentation that were tested are shown in Appendix \_. These data augmentations alter the data in a way that emphasizes useful features or deemphasizes not very useful features to help neural network to distinguish and classify these images into categories, and whether the augmentation helped with classification or not is shown at the end by its performance, in this case the accuracy of the model. Depending on which augmentation you chose and to what degree it has altered the data, it could negatively impact the accuracy also. In case of our model, it increased the accuracy and lowered its false negative rates. The classification report and confusion matrix give us more insight of this model’s performance.

Chart

Description automatically generated with low confidence

Figure 3: Proposed model with data augmentation classification report/confusion matrix

As it is shown in figure 1 and 3, the proposed model shows a great accuracy and a stable validation accuracy over numbers of epochs. More importantly, among other models with great accuracy, it showed the lowest false negative cases, where the actual cell was parasitized cell, but the model has predicted uninfected. As mentioned above, this is an important parameter of this project, as false negative cases for a Malaria patient will leave them untreated for long periods of time, where the treatment itself can even be fatal to patient. The result from other models that were used to compare the performances are shown in Appendix.

Recommendations for implementation

The traditional diagnosis of malaria in the laboratory requires careful inspection by an experienced professional to discriminate between healthy(uninfected) and infected(parasitized) red blood cells. Now, our proposed CNN model with data augmentation should be able to serve as a great automated classification system that will save tremendous time of high-paid experienced professional’s valuable time that has been consumed by tedious and time-consuming process, which can now be used for other more human touch demanding important tasks. Also, application of this automated classification technique will show higher accuracy than manual classification, as diagnostic accuracy will no longer depend on human expertise or be adversely impacted by inter-observer variability. It also eliminates the chances of human error such as fatigue or mislabeling and can run classification tasks on thousands of cell pictures while the expert may be on rest during the night or holiday.

Implementing the system would require a supervision and validation of experienced professional after classification has been complete. No system is perfect, and it will produce errors of false negatives and false positives. There needs to be a validating system where medical expert or another model goes over the images or cases that were classified as negative to detect any false negative cases. Experienced medical expert needs to put more emphasis on finding false negative cases than false positive cases, since leaving Malaria patient untreated for long periods of time could lead to situation where seeking treatment can even be fatal. Important thing to note is that when medical expert is going over images again, the images need to be the original image before the data augmentation, as data augmentation will alter images as it is shown in figure 2.

If the budget for the project allows, while this proposed model is implemented and saving time and money spent on experienced medical professional going through tedious and time-consuming task, it is also recommended to go through process of R&D as more model exploration could achieve higher model performance, meaning higher accuracy of the model or close to 0 false negative cases. As mentioned in previous paragraph, further evaluation is needed on validating system incorporating human effort or another model on finding false negative cases alongside of R&D of the model itself. If current model + human effort into finding false negative cases has accomplished almost close to 0 false negative cases determined, it may not be necessary to invest more time and money into making the current model performance close to perfect.

Appendix

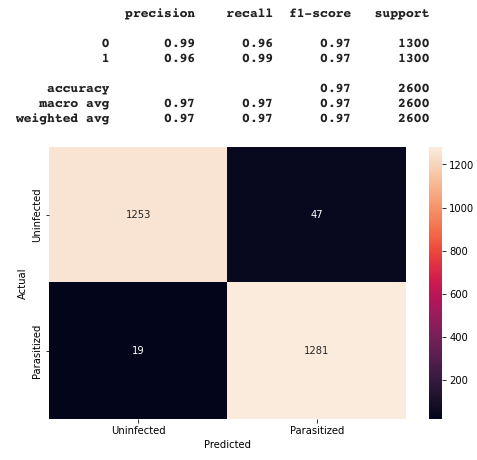
Appendix 1: Base Model performance

A picture containing chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedAppendix 2: Advanced Base Model performance

Chart, line chart

Description automatically generatedChart, treemap chart

Description automatically generatedAppendix 3: Model utilizing Batch Normalization performance

Chart, treemap chart

Description automatically generatedChart, line chart

Description automatically generatedAppendix 4: Pre-trained Model VGG 166 performance

Appendix 4: Visualization of different types of data augmentation that were tested

A picture containing text

Description automatically generated

