Diagnosing Malaria: A Comparison of Transfer Learning Approaches on the NIH Malaria Dataset with Grad-CAM Visualization for Interpretable AI

Elombe Calvert, Michelle Theodory, Ann-Marcia Tukpah, Christoph Wippel, Jason Xiang

HARVARD MEDICAL SCHOOL

INTRODUCTION

Background of Malaria

Malaria is a widespread disease caused by parasites that are carried by infected mosquitoes, causing approximately 600,000 deaths per year predominantly among children (World Health Organization, 2022). Although treatable, drug resistance and poor diagnostics make it challenging to reduce mortality rates effectively. More specifically, the most common diagnostic tool is light microscopy of blood films, which entail manual counting of parasites. The manual nature of this method is not only burdensome, but is prone to error and incorrect diagnoses. There is, then, a need for a standardised and automated method which is able to correctly identify and count parasites, and thus combat the adverse effects of Malaria.

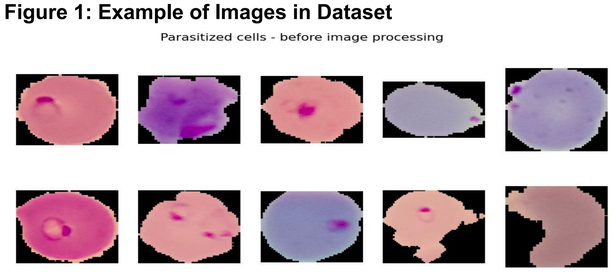
Dataset

The Malaria Cell Images Dataset retrieved from kaggle.com on March 31, 2023, which consists of 27,558 images of parasitized and non parasitized red blood cells microscopy coloured images.

Research Questions

Which CNN provides the most accurate classification of cells infected with the malaria parasite (Plasmodium) and can the architecture be improved to increase selected metrics? Specifically, we would like to prioritize improvement of the AUPRC - area under precision-recall curve (both sensitivity and PPV) because identifying cells with malaria infection accurately would be more clinically critical than identifying those without. Secondarily, we were interested in the features the model prioritizes for classification. We would examine the explainability of the above models using the GradCAM saliency method. After embarking on the project, we will likely focus on VGG19 while working through building our own CNN.

Figure 1: Example of Images in Dataset



References

1. World Health Organization. (2022). World Malaria Report 2022. Global Malaria Programme. Retrieved from https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2022.
2. National Library of Medicine (NLM). "Malaria Screener." Lister Hill National Center for Biomedical Communications (LHNCBC), https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-screener.html.

METHODS

Convolutional Neural Network

Data preprocessing:

We loaded the data using the Kaggle download API. The images had different image-sizes, therefore we resized them to all have a uniform size of 64x64 pixels and normalized the pixel values by dividing all pixels by the max value of 255. The resulting X data array for our CNN models had a shape of (27,558, 64, 64, 3). Further, we created an array of the labels (uninfected, parasitized) and stored them in a 1-dimensional numpy array of length 27,558.

Deep learning approaches attempted:

Random Forest classification (building a base model): In creating a baseline model as a reference for our ML project, we split the dataset into 80% training and 20% testing. To format our data accordingly for the random forest model, we created flattened image sequences and stored them as numpy arrays with 4,096 elements each. After that, we then trained a random forest classifier on the pre-processed data using 100 trees on the training data.

Transfer learning: The first step taken was to import the VGG19 model with pre-trained weights on the ImageNet dataset, setting the input shape to (64,64,3). The layers of the VGG19 model are then made untrainable. A few fully connected layers are added on top of the pre-trained VGG19 layers, including a dense layer with 64 units, a dropout layer with a rate of 0.5, a batch normalization layer, a ReLU activation layer, and a flatten layer. Finally, a dense output layer with sigmoid activation is added to classify the input images as either infected or uninfected with malaria. The model is then compiled with the Adam optimizer, binary cross-entropy loss, and accuracy and AUC metrics. The model is trained for 5 epochs using a train data generator and a validation data generator. These steps were also done using dataset specific random weights instead of ImageNet weights.

RESULTS

Figure 2: ROC Curve for Random Forest Model

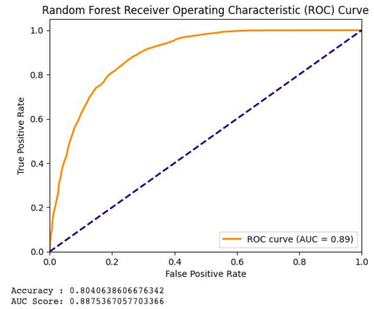


Figure 3: Accuracy - with imagenet weights vs dataset specific random weights. Blue- training accuracy, orange - validation accuracy

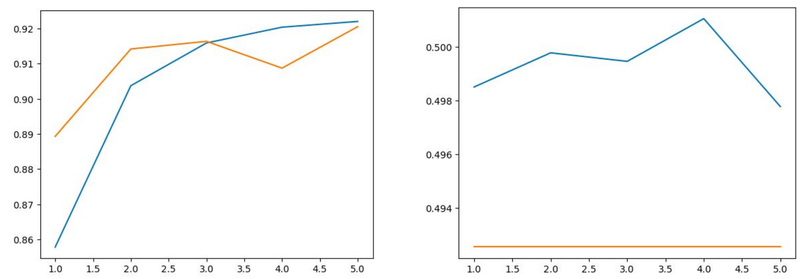
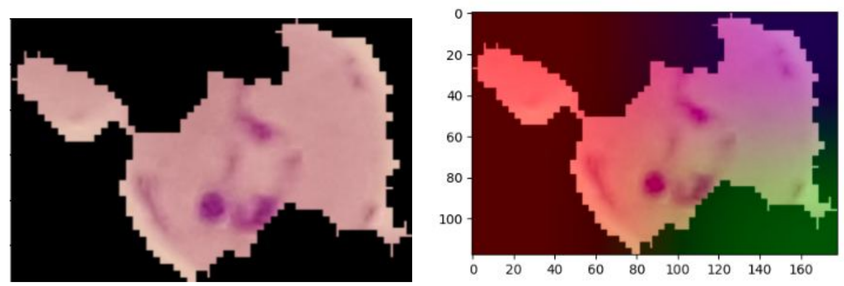


Figure 4: GradCAM Saliency Representation of Parasitized Image



DISCUSSION AND CONCLUSIONS

We found that the RF model had 80% accuracy, which was better than VGG19 with random weights. The best performing model was VGG19 using imagenet weights, which we expected. Using GradCAM, it appears the model is focusing on the right areas to predict infected cells vs uninfected cells.