## **Capstone Proposal**

Zahid Amadxarif  
April 2nd, 2019

## **Proposal**

**Domain Background**

This project aims to detect malaria using cell images from the NIH website. Malaria is a disease that is caused by parasites and is transmitted through bites of mosquitoes. It can be difficult to diagnose malaria, especially in western countries, as the health workers lack experience with it. It can be diagnosed through examining a patient’s blood under a microscope. However, identification relies on the lab technician’s experience and the quality of the microscope.

If we can support health workers by diagnosing it through images, it could help prevent many people from contracting severe malaria.

Previous research in this field has been carried out in the following paper by Antani et al (2018):

<https://peerj.com/articles/4568/>

**Problem Statement**

Identify cell images that display signs of malaria and identify cell images that do not display signs of malaria.

**Datasets and Inputs**

The dataset has been obtained from Kaggle which was donated by the NIH. It is available at:

<https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria#cell_images.zip>

<https://ceb.nlm.nih.gov/repositories/malaria-datasets/>

The dataset has 27,558 images where 13,779 of images are of infected cells and 13,779 number of images are of uninfected images.

**Solution Statement**

The most common solution to such a problem is the use of a convolutional neural network. We can experiment with different number of convolution and fully connected layers and add other layers and use different optimizers to find a model with the best accuracy.

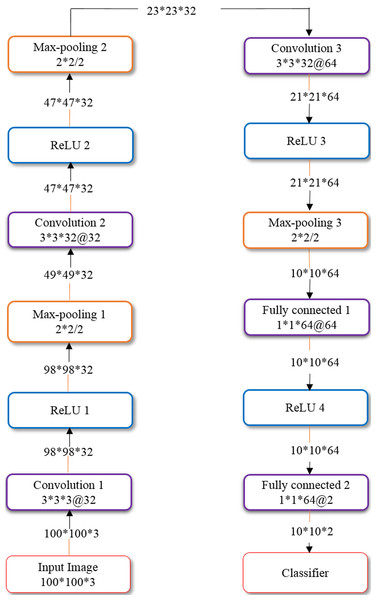
We could also apply transfer learning through utilising previously trained neural networks such as:

* AlexNet
* VGG-16
* ResNet-50

**Benchmark Model**

Antani et al (2018) utilised a customised CNN with three convolutional layers and two fully connected layers. They also carried out transfer learning using pre-trained models such as AlexNet, VGG-16, ResNet-50, Xception, DenseNet-121.

Their customised architecture is illustrated by the image below:



They found that the pre-trained ResNet-50 model outperformed the other models by achieving 0.947 ± 0.015 sensitivity and 0.972 ± 0.10 specificity.

**Evaluation Metrics**

Some common evaluation metrics include:

* Accuracy – The number of correct predictions out of all predictions made
* Sensitivity – The proportion of actual positives that are correctly identified (Number of true positives / Total number of sick individuals in a population)
* Specificity – The proportion of actual negatives that are correctly identified

(Number of true negatives / Total number of well individuals in a population)

**Project Design**

The general sequence of steps are as follows:

1. Data importing and exploring – Import the images of infected cells and images of non-infected cells
2. Data pre-processing – Add target labels for infected and non-infected cells. Split the data into training, testing and validation sets.
3. Model selection:
   1. Experiment with a CNN from scratch
   2. Experiment with a CNN using transfer learning
4. Model tuning – Fine tune the selected algorithm to increase performance without overfitting.
5. Testing – Test the model on the testing dataset.

**Notes:**

* The capstone research should compare the results found in the other paper.
* Did the ResNet-50 model also provide the best performance?
* Did our customised model do better than theirs?