# NDEV84212

Bsc Hons Data Science: Project Presentation

4-October-2019

**Olatomiwa O. Akinlaja**

SOL PLAATJE
UNIVERSITY

# Problem Statement

- About 90% of all malaria deaths in the world today occur in Africa south of the Sahara.
- The majority of infections in Africa are caused by Plasmodium falciparum.
- Identifying parasite species and their stage is very important in scutinizing the properties of malaria
- The process is labor intensive and time consuming

SOL PLAATJE
UNIVERSITY

## Why is it a data science problem

- The problem requires a machine to diagnose a disease based on microscope images of bacilli.
- A data science problem is a problem that involves data miing, cleaning and predictive modelling while also providing insights, recommendations and classifications in the process.

## Why the solution requires the skills of a data scientist

- The end result is to have a working model that has been trained to accurately recognise different bacilli.
- The process involves data gymnastics, and flexibility through complex matrix computations and manipulations.

SOL PLAATJE UNIVERSITY

# Role of Maths

The convolution of two functions, $f(t)$ and $g(t)$, is given by:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$

In discrete time, this is given by:

$$(f * g)(n) = \sum_{m=-\infty}^{\infty} f(m)g(n-m)$$

Note, however, that in general CNNs don't use *convolution*, but instead use *cross-correlation*. Colloquially, instead of "flip-and-drag," CNNs just "drag." For real-valued functions, cross-correlation is defined by:

$$(f \star g)(n) = \sum_{m=-\infty}^{\infty} f(m)g(n+m)$$

We'll follow the field's convention and call this operation convolution.

SOL PLAATJE
UNIVERSITY

The 2D convolution (formally: cross-correlation) is given by:

$$(f * g)(i,j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(m,n)g(i+m, j+n)$$

This generalizes to higher dimensions as well. Note also: these “convolutions” are not commutative.

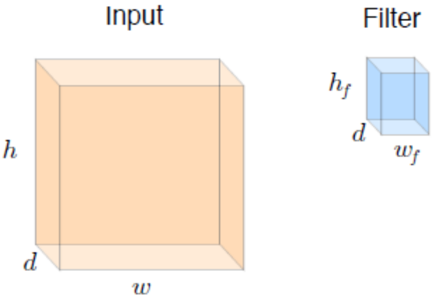
Example: Note, we assume that outside of each grid are zero values (that are not drawn). Now, dragging across the top row, we get:

$$\begin{array}{|c|c|c|}\hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array} * \begin{array}{|c|c|}\hline w & x \\ \hline y & z \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline z & y+z & y+z & y \\ \hline & & & \\ \hline & & & \\ \hline & & & \\ \hline \end{array}$$

# Convolutional Layer

This convolution operation (typically in 3D, since images often come with a width and height as well as *depth* for the R, G, B channels) defines the "convolutional layer" of a CNN. The convolutional layer defines a collection of filters (or activation maps), each with the same dimension as the input.

- Say the input was of dimensionality $(w, h, d)$.
- Say the filter dimensionality is $(w_f, h_f, d)$. So that the filter operates on a small region of the input, typically $w_f < w$.
- The depths being equal means that the output of this convolution operation is 2D.



# Data Acquisition

## Data Sources

- The data was acquired from **AI research**: "air.ug/microscopy/"
- It is structured since the data consists of images and annotations.
- The data classifies as big data since each category of disease consists of over 2000 images, each image consisting of multiple bounding boxes in order to indicate parasites and bacilli.

SOL PLAATJE UNIVERSITY

# Data Architecture

A one-off customized adapter for any camera and microscope combination is created in order to capture the images from the microscope.

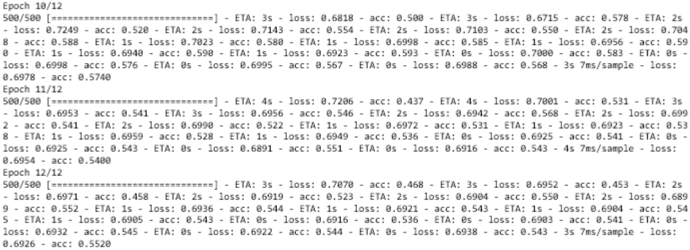
*Image capture in progress at Mulago National Referral Hospital*

The captured images then preprocessed, before being fed to a model, annotations need to be made.

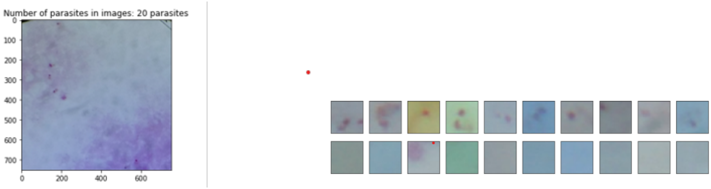
SOL PLAATJE UNIVERSITY

# Data Analysis

## Model Performance

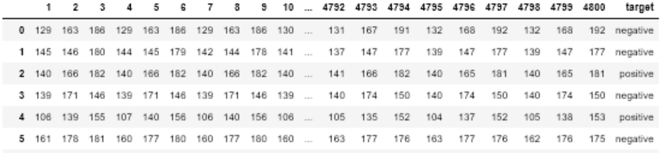


```
Epoch 10/12
500/500 [==============================] - ETA: 3s - loss: 0.6818 - acc: 0.500 - ETA: 3s - loss: 0.6715 - acc: 0.578 - ETA: 2s
- loss: 0.7249 - acc: 0.520 - ETA: 2s - loss: 0.7143 - acc: 0.554 - ETA: 2s - loss: 0.7103 - acc: 0.550 - ETA: 2s - loss: 0.704
8 - acc: 0.588 - ETA: 1s - loss: 0.7023 - acc: 0.580 - ETA: 1s - loss: 0.6998 - acc: 0.585 - ETA: 1s - loss: 0.6956 - acc: 0.59
0 - ETA: 1s - loss: 0.6940 - acc: 0.590 - ETA: 1s - loss: 0.6923 - acc: 0.593 - ETA: 0s - loss: 0.7000 - acc: 0.583 - ETA: 0s -
loss: 0.6998 - acc: 0.576 - ETA: 0s - loss: 0.6995 - acc: 0.567 - ETA: 0s - loss: 0.6988 - acc: 0.568 - 3s 7ms/sample - loss:
0.6978 - acc: 0.5740
Epoch 11/12
500/500 [==============================] - ETA: 4s - loss: 0.7206 - acc: 0.437 - ETA: 4s - loss: 0.7001 - acc: 0.531 - ETA: 3s
- loss: 0.6953 - acc: 0.541 - ETA: 3s - loss: 0.6956 - acc: 0.546 - ETA: 2s - loss: 0.6942 - acc: 0.568 - ETA: 2s - loss: 0.699
2 - acc: 0.541 - ETA: 2s - loss: 0.6990 - acc: 0.522 - ETA: 1s - loss: 0.6972 - acc: 0.531 - ETA: 1s - loss: 0.6923 - acc: 0.53
8 - ETA: 1s - loss: 0.6959 - acc: 0.528 - ETA: 1s - loss: 0.6949 - acc: 0.536 - ETA: 0s - loss: 0.6925 - acc: 0.541 - ETA: 0s -
loss: 0.6925 - acc: 0.543 - ETA: 0s - loss: 0.6891 - acc: 0.551 - ETA: 0s - loss: 0.6916 - acc: 0.543 - 4s 7ms/sample - loss:
0.6954 - acc: 0.5400
Epoch 12/12
500/500 [==============================] - ETA: 3s - loss: 0.7070 - acc: 0.468 - ETA: 3s - loss: 0.6952 - acc: 0.453 - ETA: 2s
- loss: 0.6971 - acc: 0.458 - ETA: 2s - loss: 0.6919 - acc: 0.523 - ETA: 2s - loss: 0.6904 - acc: 0.550 - ETA: 2s - loss: 0.689
9 - acc: 0.552 - ETA: 1s - loss: 0.6936 - acc: 0.544 - ETA: 1s - loss: 0.6921 - acc: 0.543 - ETA: 1s - loss: 0.6904 - acc: 0.54
5 - ETA: 1s - loss: 0.6905 - acc: 0.543 - ETA: 0s - loss: 0.6916 - acc: 0.536 - ETA: 0s - loss: 0.6903 - acc: 0.541 - ETA: 0s -
loss: 0.6932 - acc: 0.545 - ETA: 0s - loss: 0.6922 - acc: 0.544 - ETA: 0s - loss: 0.6938 - acc: 0.543 - 3s 7ms/sample - loss:
0.6926 - acc: 0.5520
```



**(a)** Parasites

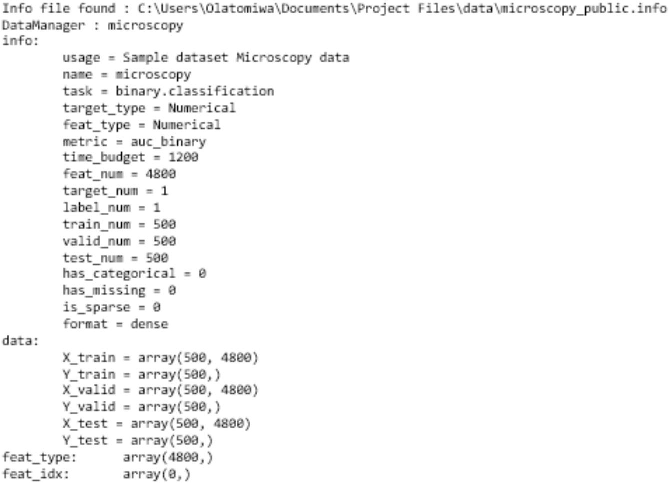
## The Data



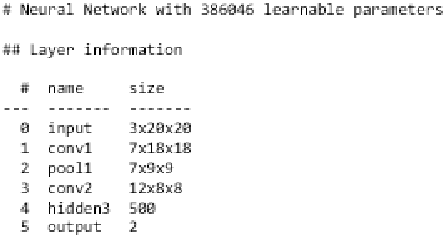
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | ... | 4792 | 4793 | 4794 | 4795 | 4796 | 4797 | 4798 | 4799 | 4800 | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 129 | 163 | 186 | 129 | 163 | 186 | 129 | 163 | 186 | 130 | ... | 131 | 167 | 191 | 132 | 168 | 192 | 132 | 168 | 192 | negative |
| 1 | 145 | 146 | 180 | 144 | 145 | 179 | 142 | 144 | 178 | 141 | ... | 137 | 147 | 177 | 139 | 147 | 177 | 139 | 147 | 177 | negative |
| 2 | 140 | 166 | 182 | 140 | 166 | 182 | 140 | 166 | 182 | 140 | ... | 141 | 166 | 182 | 140 | 165 | 181 | 140 | 165 | 181 | positive |
| 3 | 139 | 171 | 146 | 139 | 171 | 146 | 139 | 171 | 146 | 139 | ... | 140 | 174 | 150 | 140 | 174 | 150 | 140 | 174 | 150 | negative |
| 4 | 106 | 139 | 155 | 107 | 140 | 156 | 106 | 140 | 156 | 106 | ... | 105 | 135 | 152 | 104 | 137 | 152 | 105 | 138 | 153 | positive |
| 5 | 161 | 178 | 181 | 160 | 177 | 180 | 160 | 177 | 180 | 160 | ... | 163 | 177 | 176 | 163 | 177 | 176 | 162 | 176 | 175 | negative |

+ The Dataset stores an array with each row corresponding to a vectorized feature for each image.
+ Converting the 2d feature vector into one or more 1D feature vectors
+ The number of rows equal the number of images.

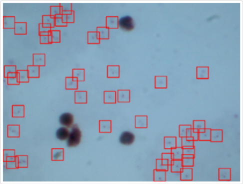
# Pre-processing



```
Info file found : C:\Users\Olatomiwa\Documents\Project Files\data\microscopy_public.info
DataManager : microscopy
info:
        usage = Sample dataset Microscopy data
        name = microscopy
        task = binary.classification
        target_type = Numerical
        feat_type = Numerical
        metric = auc_binary
        time_budget = 1200
        feat_num = 4800
        target_num = 1
        label_num = 1
        train_num = 500
        valid_num = 500
        test_num = 500
        has_categorical = 0
        has_missing = 0
        is_sparse = 0
        format = dense
data:
        X_train = array(500, 4800)
        Y_train = array(500,)
        X_valid = array(500, 4800)
        Y_valid = array(500,)
        X_test = array(500, 4800)
        Y_test = array(500,)
feat_type:      array(4800,)
feat_idx:       array(0,)
```



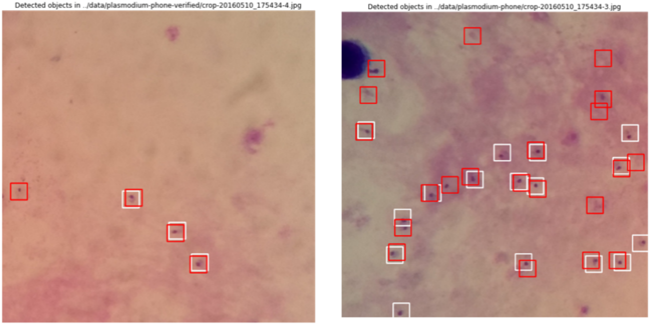
```
# Neural Network with 386046 learnable parameters

## Layer information

  #  name     size
---  -------  -------
  0  input    3x20x20
  1  conv1    7x18x18
  2  pool1    7x9x9
  3  conv2    12x8x8
  4  hidden3  500
  5  output   2
```



*Detection of plasmodium falciparum in thick blood smear image.*

**(a)** Model Parameters and the Image

# Predictions


Detected objects in ../data/plasmodium-phone-verified/crop-20160510_175434-4.jpg
Detected objects in ../data/plasmodium-phone/crop-20160510_175434-3.jpg

**(a)** Parasites Predictions

SOL PLAATJE
UNIVERSITY