

# NDEV84212

---

Bsc Hons Data Science: Project Presentation

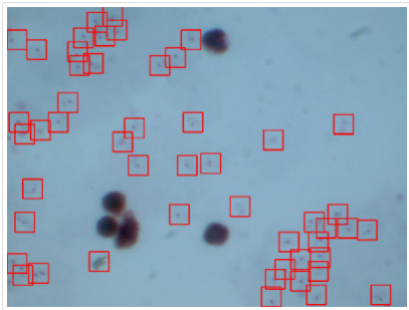
4-October-2019

**Olatomiwa O. Akinlaja**



# Problem

- + 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*.



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

## 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 mining, 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.



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.

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:

1	1	1
1	1	1
1	1	1

\*

$w$	$x$
$y$	$z$

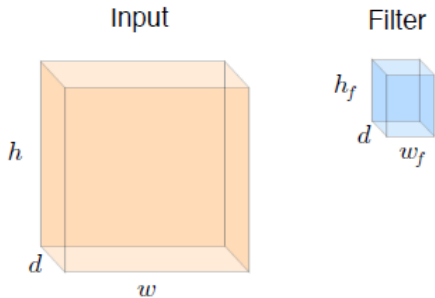
=

$z$	$y + z$	$y + z$	$y$

## 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 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 1000 images, each image consisting of thousands of bounding boxes in order to indicate parasites and bacilli.

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



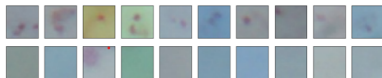
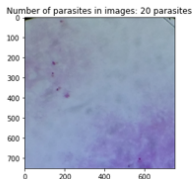
*Image capture in progress at Mulago National Referral Hospital*

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



## 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: 0.7048 - 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: 0.590 - 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: 0.6992 - 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: 0.538 - 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: 0.6899 - 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: 0.545 - 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