An Approach for Assessing Quality of Labeled Data for a Machine Learning Task in Malaria Detection

Rose Nakasi

AI and Data Science Lab, Makerere University

November 12, 2020

Malaria burden

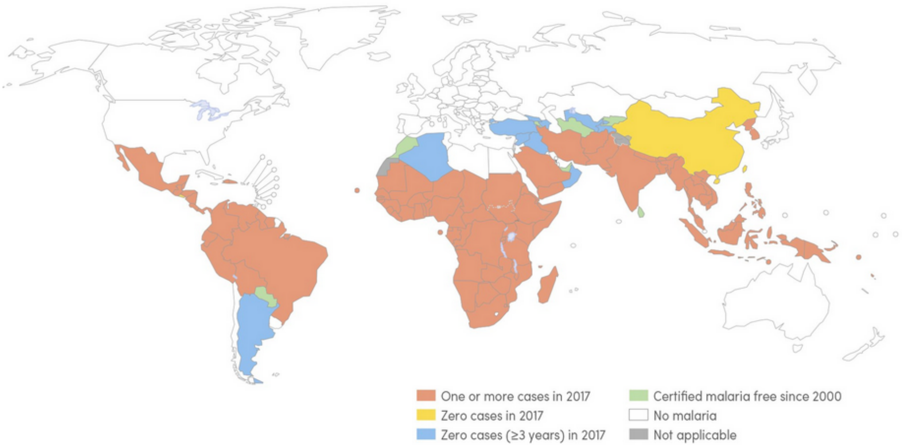


Figure: Worldwide malaria death burden(Source: WHO World Malaria Report 2018).

Motivation

-Microscopy diagnosis through supervised learning for image analysis notably contributes to malaria detection
-Manual annotation of training data is prone to inaccuracy thru;

- bias,
- Expert subjectivity
- Unclear images...

-Results into many false positives
-No study has assessed the quality of training data for malaria detection task.
-We intend to classify in respect to positives;

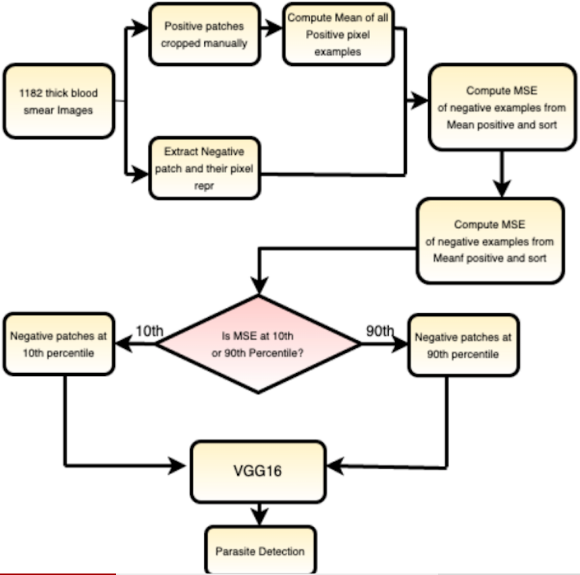
- the negative-far examples
- the negative-near examples
- To assess likelihood for false alarms in training data

Proposed methodology

We follow a six-step methodology;

- Pixel-wise extraction of patches
- Reference positive patch identification
- Class label closeness
- Threshold determination of negative examples
- Model training
- validation

Proposed methodology flow



Captured and annotated image

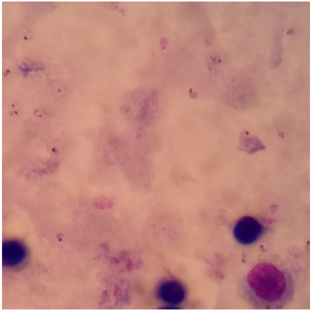


Figure: Captured image

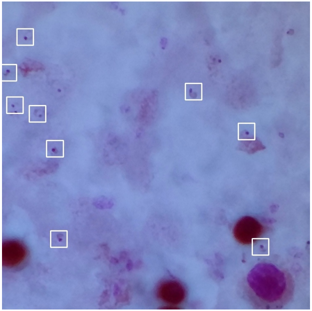


Figure: Annotated image

Step 1: Pixel-wise extraction of Patches

From 1182 thick blood smear images captured,
-Positive Patches -taken centered on bounding boxes in annotation
-Negative patches-no bounding box

Table: Positive and negative patches generated

Threshold used	Pos patches	Neg patches
At 10%	7045	147086
At 90%	7045	16730

Step 2. Identifying reference pixel-wise positive patch

Using image averaging algorithm;
Compute arithmetic mean of intensity values of each pixel position in a set of positive patches. for each positive the mean pixels for the entire positive patch examples, Calculate the mean pixels for the entire positive patch examples,

$$A(N, x, y) = \frac{1}{N} \cdot \sum_{i=1}^{N} I(i, x, y) \tag{1}$$

Step 3. Class label closeness

The MSE between two images (positive,p and negative,n) defined as p(x,y) and negative n(x,y) is defined as shown in equation,

$$MSE = \frac{1}{XY} \sum_{y=1}^{X} \sum_{x=1}^{Y} [p(y,x) - n(y,x)]^2 \tag{2}$$

Goal: To compute the different MSE values that correspond to each negative patch.

Step 4. Threshold determination of negative examples

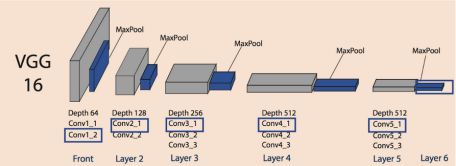
Criteria for selection of the negative samples.
Fistly, sort MSE values of negative patches(from small value to big)

- At 10^{th} percentile
- At 90^{th} percentile

Step 5. Model training/selection

to classify negative patches;

- VGG 16 Model was used
- Data was split into 60%, 20% and 20% (train, test and Val)



VGG16 has smaller filters;

-reducing the number of parameters thus increasing the non-linearity

Step 5. Results

Classification Accuracy;

Threshold used	Classification
At 10%	0.7036
At 90%	0.9126

- approach can thus aid in the selection of quality training dataset for a malaria detection task.
- transparent , accountable and trust worthy data for machine learning solutions.

Thank you!

Questions?