Evaluating the Performance of Automated Classification of Sputum Smear Slides for TB Diagnostics

Ali Akbar Septiandri[1], **Muhammad Rezqi**[1], **Dewi Nur Aisyah**[1,2], **Ahmad Ataka Awwalur Rizqi**[1,3], **Vani Virdyawan**[1,4], **Dwi Marhaendro Jati Purnomo**[1,4] **Daniek Suryaningdiah**[5], **Alfian Nur Rosyid**[6]

[1]Garuda45, [2]University College London, [3]King's College London, [4]Imperial College London, [5]Public Health Office of Surabaya City, [6]Universitas Airlangga

tbdecare@garuda45.org

1 Introduction

Some research has been conducted to detect *Mycobacterium tuberculosis* (MTB) in image patches extracted from sputum smear slide photos under a microscope by employing deep convolutional neural networks [Quinn *et al.*, 2016]. To verify that this automated TB diagnostic tool really does increase sensitivity and specificity of detection, an evaluation of the tool in real-life scenario is needed.

We implemented and evaluated the work of Quinn et al. [2016] in automated classification of sputum samples for tuberculosis (TB) diagnostics in Surabaya, Indonesia in a real-life setting. Sputum samples in this research were collected in Surabaya as it ranked second in terms of absolute numbers of TB cases and Surabaya is the capital of East Java, a province in Indonesia where TB is second most prevalent. Two sputum samples were collected each from TB suspects and follow-up patients across 30 public health centres and 1 hospital to be tested using Acid Fast Bacilli (AFB) and bacterial culture methods. Our research has shown that we can predict all the bacterial culture test results correctly by applying logistic regression to the bacteria count in image patches from photos of sputum slides. We believe that this evaluation can be a stepping stone to achieve nation-wide implementation of this model to detect TB cases effectively and efficiently.

2 Cohort

Two sputum samples (on site and early morning) were collected from TB suspects and follow-up patients across 30 public health centres and 1 hospital in Surabaya, Indonesia from February to May 2018. Sputum samples were processed for Acid Fast Bacilli (AFB) detection using Ziehl-Neelsen (ZN) staining. Stained slides were then read by a lab technician at the public health centres and by the automated TB diagnostic tool using pictures taken from the microscope. Pictures of the slides were taken based on the WHO standard of 100 fields of view at 1000$\times$ magnification. The bacterial culture tests were done by Centre for Health Laboratory of Surabaya.

3 Methods

We identified MTB bacteria using the model provided by Quinn et al. [2016] and counted the number of image patches which contain a bacterium. We sliced each ZN-stained slide photo, that represents a field of view, into 160x160 pixel image patches. Since the photos are resized into 1632x1224 pixels according to [Quinn *et al.*, 2016], we will have 10x7 image patches for every photo. We then calculate the sensitivity and specificity of this diagnostic tool by applying logistic regression to the bacteria counts from each field of view with the corresponding AFB and bacterial culture test results as the labels since it is considered the diagnostic gold standard for active TB.

4 Results

We ran 3-fold cross-validation logistic regression with imbalanced weighted classes to achieve 94.57% accuracy. The labels in this case came from standard AFB detection which consists of five classes, i.e. negative, scanty, +1, +2, and +3. However, what we really need to predict is the golden standard, i.e. bacterial culture test results.

Using the same method to predict the AFB results, we also trained a model to predict the outcome of bacterial culture results and got 100% accuracy. As a comparison, we classified samples with scanty, +1, +2, or +3 AFB results as positive and got 91.30% accuracy. The sensitivity and specificity from this method are 78.26% and 4.35% respectively. Table 1 depicts a comparison of each method's result.

However, we had to adjust the class weight to achieve those results. To make the model more sensitive, we assigned 1:5 for negative:positive weight ratio. Interestingly, even when adding even more weight to the positive class, we would not get false positives.

5 Discussion

We believe that using logistic regression has enabled us to learn the boundary between each class better than the previous standard by WHO, thus maximising the accuracy. We hope that our model can help lab technicians by automating TB detection from AFB so that the consistency of the results can be attained.

To this date, we have obtained around 545 samples, but most of them were not digitised yet. We are also waiting for the rest of the bacterial culture test results as the ground truth to our prediction.

Table 1: A comparison of each method's result

Method	#Classes	Classes	Accuracy	Sensitivity	Specificity
3-fold cross-validation logistic regression with imbalanced weighted classes	5	AFB (negative, scanty, +1, +2, and +3)	94.57%	-	-
3-fold cross-validation logistic regression with imbalanced weighted classes	2	Culture (negative, positive)	100%	100%	100%
Rule-based (scanty, +1, +2, or +3 are classified as positive)	2	Culture (negative, positive)	91.30%	78.26%	4.35%

References

[Quinn *et al.*, 2016] John A Quinn, Rose Nakasi, Pius KB Mugagga, Patrick Byanyima, William Lubega, and Alfred Andama. Deep convolutional neural networks for microscopy-based point of care diagnostics. In *Machine Learning for Healthcare Conference*, pages 271–281, 2016.