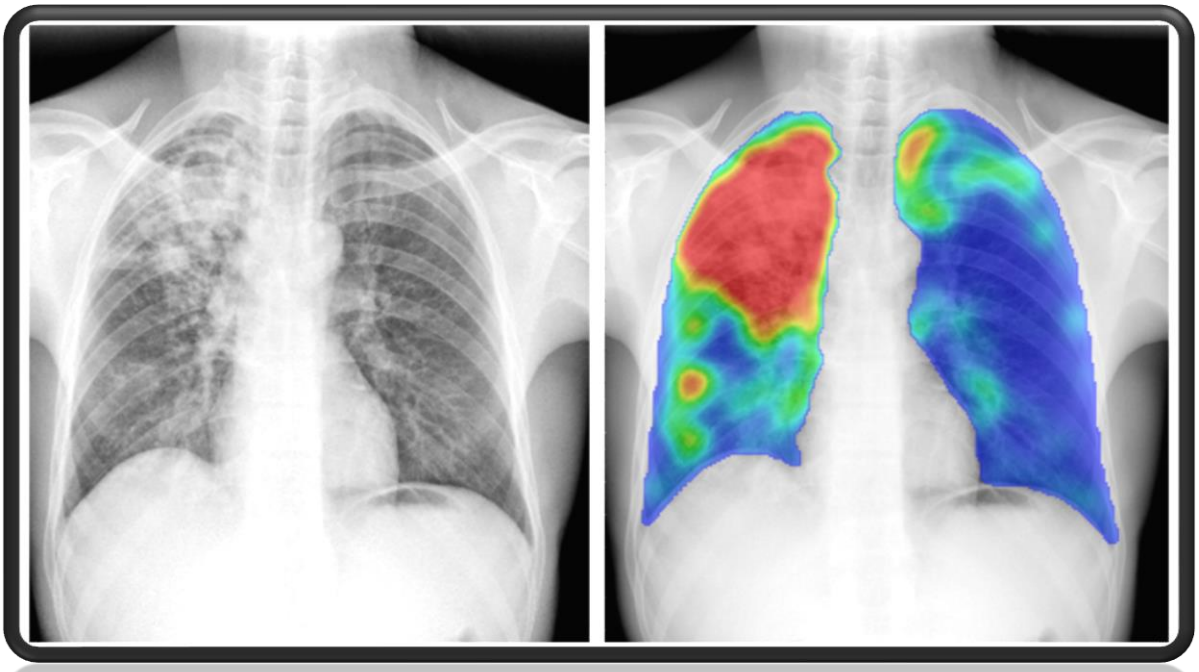


CAD4TB White paper

2018-08-23

CAD4TB 6

COMPUTER AIDED DETECTION FOR TUBERCULOSIS

ABOUT THIS WHITE PAPER

This white paper applies to CAD4TB 6 and describes the general principles of the CAD4TB software, explains matters like ROC curve, sensitivity, specificity, threshold values and how to interpret the results.

Finally, Annex A describes our Verification Procedure for CAD4TB which should be done for every new X-ray scanner - population combination.

IMPORTANT NOTICES

Delft Imaging Systems does not assume responsibility for the misuse of the CAD4TB software. Since Delft Imaging Systems cannot control the use of CAD4TB, it shall not be held responsible for any direct or consequential personal injury or damage.

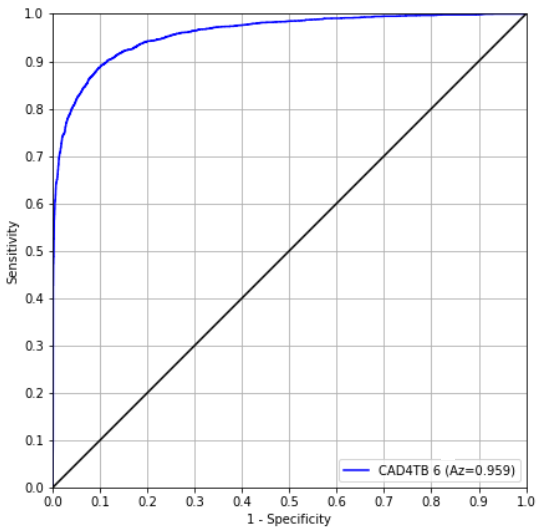
INTRODUCTION

CAD4TB is a software product that takes a single frontal chest radiograph as input, in the form of a DICOM image, and produces several outputs. The output includes a quality assessment of the input image, a heat map indicating areas with possible abnormalities, and an output score between 0 and 100 related to the likelihood that the X-ray is radiologically abnormal and the subject on the X-ray has tuberculosis (TB).

Users can take these outputs into account in their clinical work: they can decide if a new image should be acquired, in case the quality assessment indicates suboptimal image quality; they can decide that the subject should undergo further testing for the presence of TB or other lung diseases in case the heat map displays suspicious regions that are verified by a human operator as suspicious or when the score is above a certain threshold. The choice for this threshold should be made by the user and will depend on the conditions under which the software is used.

DEFINITIONS

- **Sensitivity**: proportion of positive images (i.e. having TB) correctly labeled as positive.
- **Specificity**: proportion of negative images (i.e. not having TB) correctly labeled as negative.
- **ROC curve**: this curve is created by plotting the True Positive rate (sensitivity) against the False Positive Rate (1-specificity) at various threshold settings.
- **Az**: area under the ROC curve. This number estimates the probability of correct ranking positive/negative. It is bound between 0 and 1: the closest to 1, the better the system's performance. 1 means a perfect classification.



A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

- 0.90-1 = excellent (A)
- 0.80-0.90 = good (B)
- 0.70-0.80 = fair (C)
- 0.60-0.70 = poor (D)
- 0.50-0.60 = fail (F)

- **T**: threshold value. T=50 corresponds to the point on the ROC curve closest to perfect classification, but is not necessarily the optimal threshold for the most cost-effective TB screening approach (see section *Which threshold to use*). **T** is the value set by the user to determine which images are labeled abnormal/normal.

Table 1. CAD4TB threshold values with their corresponding sensitivity and specificity levels.

Threshold	Sensitivity	(1-Specificity)
0	100	100
20	100	75
40	99	52
50	91	13
60	75	2
80	35	0
100	0	0

This can be presented graphically with an ROC curve:

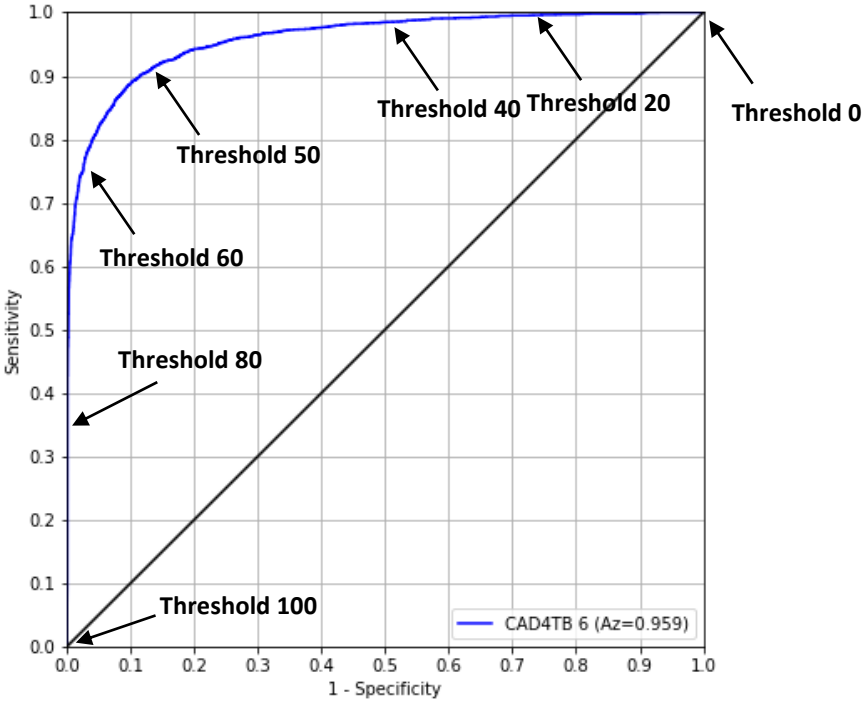


Figure 1. CAD4TB ROC curve.

To summarize: An ROC curve demonstrates several things/characteristics of the test:

- ✓ It shows the trade-off between sensitivity and specificity (any increase in sensitivity will often be accompanied by a decrease in specificity).
- ✓ The closer the curve follows the left-hand border and then the top border of the ROC space, the better the test´s performance.
- ✓ The closer the curve comes to the 45-degree diagonal of the ROC space, the worse the test's performance.
- ✓ The area under the ROC curve is a measure of the test's performance.

CAD4TB: HOW DOES IT WORK?

CAD4TB is software based on the principles of deep learning. In the process of computing its score, it compares regions in the radiograph with regions extracted from normal and abnormal radiographs. These latter radiographs form the training data set of the software. A basic principle of deep learning is that the training data should be representative of the test data, otherwise the results may not be reliable.

X-ray units from different manufacturers produce different qualities of images because of hardware differences and because manufacturers apply different post processing algorithms to their images in order to improve the image quality and equalize local contrast. In addition, scan protocols may vary. For example, the kilo voltage setting of the X-ray tube has a large effect on the relative contrast of bony structures. As a result, computer algorithms such as CAD4TB may have difficulties interpreting images from different sources. Similarly, the subjects depicted on the radiographs may originate from different populations in which the manifestations of TB vary.

Specific algorithms to normalize radiographs prior to analysis have been included in the CAD4TB software, but these algorithms have limitations. As a result, we have set up a formal test procedure to verify if data from a new X-ray unit or a new population can be reliably processed with CAD4TB (Annex A).

Currently, CAD4TB has been verified on the X-ray models listed in Table 2.

Table 2. Verified X-ray machines

MANUFACTURER	MODEL
Canon/ DelftDI	CXDI series
Agfa	CR10-X, CR12-X
Philips Medical Systems	Dura Diagnost Compact
Samsung Electronics	DGR-U6QN2B/US
Swissray Medical AG	ddR Element, ddR Chest
3DISC	QuantorMed

CAD4TB: HOW TO READ THE SCORE?

CAD4TB 6 is trained in such a way that a score of 50 corresponds to the point on the ROC curve closest to perfect classification (the upper left corner, see Introduction). This calibrated 50-score point represents the best trade-off between sensitivity and specificity. We have tested this "50" point with several certified human CXR readers and they all have their 50-score point close to this calibration point. Having this calibration based on the ROC curve results in scores for similar images being very close even if they come from different X-ray devices as can be seen in the examples in Figure 2.

Example images read as positive (score > 50)

*Read as **positive** (Canon)*
CAD4TB score: 85.2

*Read as **positive** (Agfa)*
CAD4TB score: 81.4

*Read as **positive** (Philips)*
CAD4TB score: 84.4

Example images read as negative (score < 50)

*Read as **negative** (Canon)*
CAD4TB score: 26.9

*Read as **negative** (Samsung)*
CAD4TB score: 31.1

*Read as **negative** (Swissray)*
CAD4TB score: 24.0

Figure 2. Example X-rays with CAD4TB score

CAD4TB: WHICH THRESHOLD TO USE?

As mentioned before, a threshold value of 50 corresponds to the point on the ROC curve closest to perfect classification. However, this might not be the optimal threshold for the most cost-effective TB screening approach, as cost effectiveness may be based on multiple factors, such as the laboratory capacity to do GeneXpert (GXP) tests or the budget available to perform a screening program. **In consequence, we do not select a threshold for the software but supply a table with different threshold levels and the corresponding sensitivity and specificity levels after following a verification procedure (Annex A) so users can determine the most suitable software configuration depending on their operating conditions.**

Table 3 below show different thresholds and sensitivity and specificity levels for different X-ray/detector systems to show variation among the machines. The *% ref* is the percentage of images that falls above the threshold. This percentage is computed as follows: assuming that 7% of the images are abnormal (prevalence), then 93% of the images are normal. Using a threshold of 50 (highlighted in yellow), sensitivity is 91%, which means that 91% of the abnormal images will be flagged correctly as "positive," corresponding to 91% of 7% = ~6 % of the screened subjects. On the other hand, a specificity of 87% means that 100% - 87% = 13% of all normal images will be flagged as "positive" as well, which corresponds to 13% of 93% ~12 % of the screened subjects. Accumulating the two, results in 18% and this is the percentage of cases that would be referred to e.g. a GXP MTB/RIF test.

Table 3. CAD4TB 6 thresholds and performance

	CAD4TB 6		
Threshold	***% ref cases***	**Specificity (%)**	**Sensitivity (%)**
10	*90*	11	100
15	*83*	18	100
20	*77*	25	100
25	*70*	32	99
30	*66*	37	99
35	*60*	43	99
40	*55*	48	99
45	*41*	63	97
50	*18*	87	91
55	*11*	94	84
60	*7*	98	75
65	*5*	99	65
70	*4*	100	56
75	*3*	100	45
80	*2*	100	35
85	*2*	100	26
90	*1*	100	17

ANNEX A: CAD4TB VERIFICATION PROCEDURE

The CAD4TB verification procedure consists of two mandatory stages that perform the actual verification and a third, optional stage that corresponds to reoptimizing the software in case the outcome of verification was not satisfactory.

STAGE 1

In the first stage, a set of 10 DICOM images are provided to the CAD4TB development team. It is verified if these images can be read correctly and if the software produces plausible results. Shortcomings are identified.

STAGE 2

If the first stage does not reveal problems, in a second phase we ask the source to provide at least 100, and preferably around 200, images that are considered normal and at least 100, and preferably around 200, images that are considered abnormal. It is crucial that the selection process of these cases is random, e.g. that the normal images are considered representative of the distribution of normal images from the source, and similarly, that the abnormal images are considered representative of the abnormal images from that source.

The CAD4TB team will produce a report that includes any foreseen issues and possible solutions, examples of results per image, and performance statistics in the form of ROC curves and sensitivity/specificity tables. These results will be compared with known results from other sources. The outcome of this stage is a prediction if CAD4TB will work as well on data from this source as on data from other known sources, and a recommendation to use or not use CAD4TB for the analysis of data from this source.

STAGE 3

If a recommendation to not use CAD4TB is made after the second stage, a software optimization stage could be proposed. For this stage, an additional set of 250 normal and 250 abnormal images will be asked. The same conditions related to randomness and representativeness as explained for the second stage apply to this third stage. If the source can provide these data, the CAD4TB team will include them in the training data set used by the software and proceed to retrain the pertinent components. After this, the second stage above will be repeated. It is expected that the performance of CAD4TB improves given the newly used data. If the resulting performance is appropriate, a new recommendation will be made.

REQUIREMENTS

Specific requirements for the images that are sent:

- Image format: Dicom (DCM)
- Subjects on the images should be at least 4 years old.
- Posterior anterior view
- Digital images either from a CR or DX source (scanned films are not usable/acceptable)

Information to be provided:

- Contact name and address
- Brand and type of X-ray equipment (scanner and detector)
- Description of the setting in which the X-rays were/are acquired (TB clinic, active case finding, prevalence survey, etc.)

CONTACT

For questions and setting up a data transfer procedure, please contact:

- Rick Philipsen (rickphilipsen@thirona.eu)
- Annet Meijers (annetmeijers@thirona.eu)