Deep Learning for Automated Detection of Pulmonary Tuberculosis from Lung Ultrasound

Julien Tuấn Tú Vignoud
EPFL
vignoud.julien@gmail.com

Mary-Anne Hartley
EPFL
mary-anne.hartley@epfl.ch

Abstract

Background. Despite being both preventable and curable, tuberculosis (TB) remains the world's leading cause of mortality. Existing sputum-based diagnostic solutions require expertise and logistics which limit their widespread use in low-resource settings. Lung ultrasound (LUS) is a promising alternative, being cheaper, non-invasive, virtually consumable-free and applicable at point-of-care. However, even expert interpretation is poorly specific or sensitive for the detection of TB.

Aim. We present a deep learning (DL) algorithm to automate and improve the diagnostic capacity of lung ultrasound for pulmonary TB.

Methods. Systematic LUS exams were performed on 110 adult patients suspected of lower respiratory tract infection on presentation at the Lazeret pulmonology outpatient department in Cotonou, Benin. We use the resulting 2674 LUS images to train an interpretable DL model that classifies patients into TB+ vs TB- categories based on the sputum-based molecular gold standard (GeneXpert-Ultra). External validation is performed on an independently recruited cohort of 30 patients from rural South Africa. We evaluate the relative diagnostic importance of anatomic positions from which ultrasound images are acquired. Grad-CAM further highlights the regions of each image the model finds most determinant. The highlighted regions were then qualitatively assessed by human LUS experts in order to evaluate the model's clinical plausibility by its alignment with physiology and pathology. Finally, results are compared to a "human baseline" model, which inputs clinician-engineered features.

Findings. The DL model performs significantly better than the model based on clinician-extracted features ($p < 0.001$) with a area under the receiver-operator curve (AUROC) of 0.88 (95% CI $[0.86, 0.91]$) vs 0.57 (95% CI $[0.56, 0.59]$). Moreover, LUS experts assess that anatomic locations used by the model are relevant and we find that LUS experts and the model make the same number of incorrect diagnoses. Finally, we find that the DL model performance remains stable when using only 3 lung locations compared to all 14 possible.

Conclusion. Automated DL diagnosis from LUS on this population has potential for TB triage, meeting the WHO criteria for such a tool with 90% sensitivity and 70% specificity. Further work is required on larger cohorts to improve the generalisability necessary for real world deployment.

1 Background

1.1 Tuberculosis

While the COVID-19 pandemic had an enormous direct health impact, its indirect disruption of access to tuberculosis (TB) services was arguably as significant. In 2020 the WHO reported a global incidence of 10 million cases with 1.3 million deaths: 100'000 more than anticipated, thus marking the first annual increase in TB-related deaths since 2005. [34].

Among countries with prevalence surveys, 30 are estimated to contribute to 86% of the global prevalence. Two thirds of this is concentrated in just eight countries, of which South Africa is one, where the proportion of people with TB reaches 900 per 100'000 [24]. Western Europe, by contrast, experiences less than 1 death per 100 000 population per year. Indeed, TB is a disease of poverty. Data from 21 countries shows that TB in turn further compromises financial security: affected households in these countries spend at least a fifth of their income on TB-related costs [4] [2]. It is the single largest infectious cause of death, which is particularly remarkable, given that TB is both preventable and treatable.

1.2 Existing solutions for TB diagnosis

Diagnostic tests. The current methods used for diagnosing TB may have excellent performance in ideal conditions, but require expertise and logistical organisation that limit their widespread use in low-resource settings.

The current gold standard for TB diagnosis recommended by the Centers for Disease Control and Prevention is sputum-based microbiology [11]. However, cultures may take between 4 and 12 days and require rigorous laboratory expertise and logistics to ensure sample quality and that patients are not lost to follow-up. Similarly, the nucleic acid amplification geneXpert tests depend upon investments beyond the scope of most facilities where screening could potentially take place. While the WHO recommendation is to use these methods to replace the Tuberculin skin test (TST, which performs sub-optimally in BCG-vaccinated populations [32]) and the sputum smear test (which lacks sensitivity [14]) many countries still employ these outdated techniques [10] due to the difficultly and expense of the new recommendations.

Clinical diagnosis. The above tests are usually performed based on clinical and epidemiological suspicion, where key symptoms (weight loss, prolonged cough, night sweats etc.) and risk exposure (HIV, TB contact) create a clinical case definition warranting further testing. A chest X-ray can add a significant amount of sensitivity and specificity but is unable to distinguish between past TB scarring and new active disease. Further, it exposes patients to ionizing radiation and requires patient mobility which is unfeasible for severe cases.

1.3 Point-of-care ultrasound.

In contrast, point-of-care ultrasound is a non-ionizing, low-cost and virtually consumable-free tool which can be deployed at the point-of-care. New ultrasound on a chip technology has made these tools portable and affordable to remote and low-resource settings. Another benefit of using ultrasound is the low risk of cross-infection when using a plastic disposable cover on a portable handheld device.

Images are generated in real time by transmitting sound into the tissues by a "transducer" and then recording those that are reflected back with a "receiver". The time taken for the wave to return determines depth (rendered as distance on the y axis of the image), while the number of waves received determines their echogenicity (rendered as brightness). As the presence of an air/tissue interface scatters ultrasound, nothing is visible past air bubbles. This fact was originally thought to make lung ultrasound unfeasible. However, it became increasingly appreciated that the presence of liquid and consolidations at the pleural line made patterns that had high sensitivity and specificity for the diagnosis of pneumonia, infiltrates and other disease. Until now, however, the infectious aetiology of the pneumonia could not be determined by ultrasound alone.

See Figure 1A for an overview of this ultrasound image and the acquisition protocol for LUS involving capturing 14 anatomic sites.

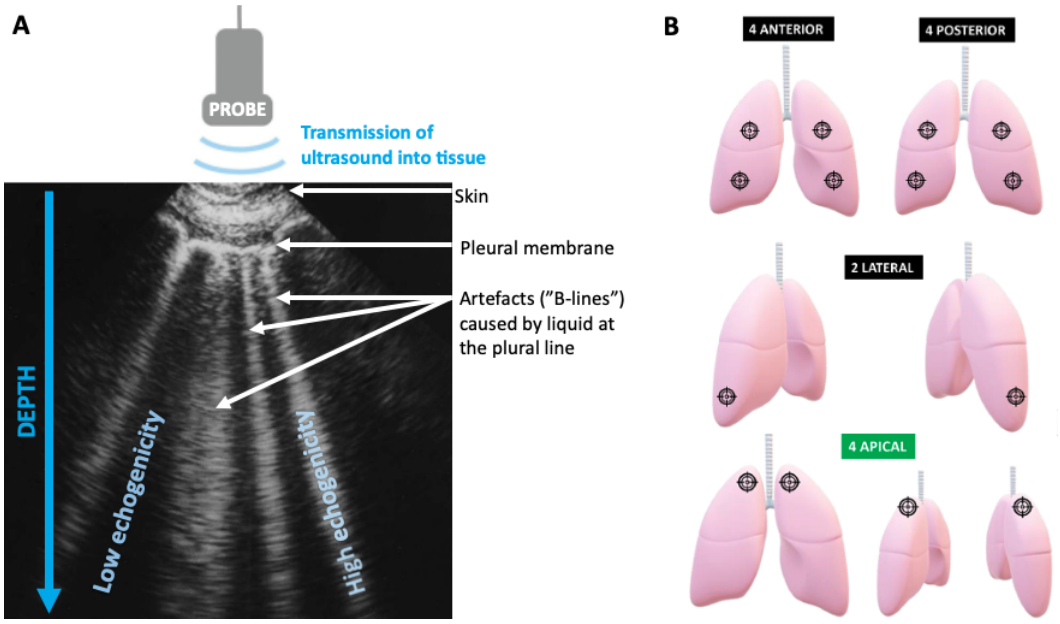


Figure 1: **(A)** Overview of the main elements in a lung ultrasound image. **(B)** The acquisition protocol of LUS, involving the collection of images from 14 anatomic locations on a single patient

Ultrasound also presents challenges like high inter-operator variability, as the image acquisition and evaluation demand expertise. For example, a study on the prenatal detection of malformations using US images demonstrated that the sensitivity ranged from 27.5% to 96% among different medical institutes[26]. This variability remains a serious challenge in ultrasound-based clinical decision making. Additionally, as can be seen in Figure 1B, the acquisition protocol requires 14 images, which is time consuming.

Computer assisted diagnostics which leverage deep learning (DL) could automate interpretation, with the added potential to detect patterns not visible by the human eye. Such tools have the potential to decrease variability of predictions [9] [17]. In addition to minimizing operator error, automated detection affords the possibility of rapid processing of a vast amount of data as well as minimizing the information needed per patient, thus optimizing the acquisition protocol. A robust automated system might even be able to guide clinical decision making in scenarios where resources are scarce and trained personnel are unavailable. Deep Learning has recently made major advances in the field of medical image analysis [29] [23]. For ultrasound, a review of the application of DL in muscle imaging underscores the need for more robust methods for image interpretation and advises the use of DL algorithms to overcome the differences between manufacturers' devices.[33]

2 Aim and objectives

In line with WHO's recommendation to expand the use of digital technologies [3] and an increased need for efficient evaluation of ultrasound images, we develop a new deep learning method to improve the diagnostic capacity of LUS for TB.

2.1 Objectives

1. **Literature review** To review the literature for existing methods, guidelines and pitfalls.
2. **Preprocessing pipeline** To preprocess the data and screen for errors, duplicates and potential sources of bias.
3. **DL diagnostic algorithm** To develop a model that predicts the geneXpert TB diagnosis of a patient using only LUS images as input.

4. **Human baseline algorithm** To develop a model that predicts the geneXpert TB diagnosis of a patient using clinician handcrafted features (i.e. tabular data recording the binary presence/absence of human-visible patterns in ultrasound).
5. **Clinical plausibility of the DL algorithm** To evaluate the alignment of the model's predictions with human expert analysis using interpretability techniques
6. **Optimize acquisition** To compare the importance given to the different anatomic sites to find the optimal combination and number images required for prediction.

3 Related work

DL in TB diagnosis. To the best of our knowledge, no AI method using ultrasound images has been developed to diagnose TB. However, existing works make use of AI for TB diagnosis with different inputs. For instance, Dande et al. [12] review the use of Artificial Neural Networks (ANNs) as a diagnostic tool, on which most of the Deep Learning field relies. ANNs draw inspiration from our brain or the biological neural networking system. The first use of ANNs for TB diagnosis was carried out in 1999 and aimed at predicting the prevalence of TB from tabular data such as radiographic findings, symptoms and demographics. The trained model showed a sensitivity of 100% and a specificity of 72%. Following studies leveraging ANNs reached similar results. Thus, it was derived that neural networks were a potential diagnostic tool for tuberculosis.

More specifically, Kulkarni et al.[21] analyze the use of DL algorithms to predict TB from chest radiography. The first DL model for TB detection was created in 2016 by Hwang et al. [20]. They used a model called AlexNet that directly takes images as inputs. The model was trained on 10'800 chest x-rays and reached an AUROC between 0.88 and 0.964 on their test sets. Lakhani and Sundaram [22] found that their most accurate approach utilized an ensemble of models together with a radiologist to adjudicate discrepant cases, which achieved a sensitivity of 97.3%, specificity of 100%, and AUROC of 0.99. They suggest that the best use of such algorithms may be to augment to capabilities of radiologists working in resource-poor regions.

DL in ultrasound. None of the previous works makes use of lung ultrasound, perhaps because of the poor predictive capacity when interpreted by humans. Brattain et al. [9] review ML applications for medical ultrasound in general and shows that DL approaches can significantly improve performance when compared with human interpretation or classifiers operating on handcrafted features. Furthermore, they address the need for results to be interpretable by clinicians, which is not systematic in the previous works. Akkus et al. [5] survey more precisely DL applications of medical ultrasounds. They assess that the generalization ability of DL-based diagnosis approaches is superior than traditional ML approaches. Though, it is still hindered by the variability in ultrasound images. They suggest the use of transfer learning, i.e., pre-training models on large scale image datasets, as well as data augmentation (e.g. translation, flips, distortion of images) to increase the robustness and generalization of models.

DL for LUS. While DL has not used ultrasounds to predict TB, extensive resources exist for COVID-19. Zhao and Bell [35] review such DL models for COVID detection. Among the methods studied, 6 "out-of-the-box" architectures are used and 3 new architectures are presented to improve their robustness and generalization. Notably, Born et al. [8] employ class activation maps (CAM) to improve prediction interpretability. CAM overlays the original ultrasound images with the pixel-wise importance in the model prediction. Figure 11b illustrates such CAM. The best performing diagnostic model achieved 94.39% accuracy, 82% precision, 76% sensitivity, and 96% specificity [19] by leveraging a ResNet architecture [18]. The review establishes that models lack training on balanced dataset and while interpretable methods exist, they are still not used systematically.

4 Methods

4.1 Datasets

The cohort comprises 111 adult patients recruited as part of a prospective observational study on the diagnostic potential of lung ultrasound in TB-endemic regions. Inclusion criteria were consenting adults with cough and/or dysnopea of any duration. However, patients were excluded if cough came from a definite non-infectious origin such as asthma or heart failure. Patients were recruited on presentation at the Lazeret pulmonology outpatient department in Cotonou, Benin between October 2021 and March 2022. Systematic LUS exams were performed following the protocol depicted in Figure 1B resulting in an image bank of 2674 LUS images. The labels for TB+ vs TB- categories are derived from sputum-based molecular gold standard (GeneXpert-Ultra). Of the 111 patients, 40 (36%) are TB+ by sputum-based GeneXpert-ultra and 71 (64%) negative. External validation is performed on an independently recruited cohort of 35 patients from Tintswalo Hospital in Mpumalanga South Africa. 6/35 (9%)patients are TB positive. Several negative cases are expected to be false negatives due to pauci-bacillary yield in HIV+ patients. Benin is considered to have a moderate-to-low endemicity for HIV and TB, while South Africa has one of the highest incidence rates in the world of both diseases.

4.2 Literature review

Existing works were first reviewed on PubMed [1] as a source of peer-reviewed medical papers. As we have seen, the specific combination of DL for TB using LUS is unprecedented. Therefore, the search was extended to larger scopes. The search on PubMed was conducted using MeSH terms, controlled keywords with defined and specific significations, combined with the Boolean operator AND. Here is a list of the different searches:

- *Artificial intelligence* AND *tuberculosis*
- *Artificial intelligence* AND *ultrasonography*
- *Artificial intelligence* AND *ultrasonography* AND *SARS-CoV-2*

A similar search was conducted in the Cochrane Library [2], which references systemic reviews in health care and health policy. The only work found was a review of thoracic imaging tests for COVID-19 [16]. Subsequently, medRxiv [3], a database of medical preprints, and Google Scholar were searched despite containing unreviewed works for the sake of completeness.

4.3 Data preprocessing

The original data is located on the Butterfly Cloud, where images collected by ultrasound probes are directly uploaded. The first challenge arises in retrieving images and videos from the online platform, which used to be done manually, downloading each of them one by one. To address this issue, we implemented a user-simulating software that fetches all the files in an automated manner. A repetitive task lasting for a day or two now takes one hour. The tool is public and free of use [4].

Refer to Figure 12a for an example of a LUS image. The images are manually inspected for unusable images, such as corrupted or zoomed-in images. Unless specified otherwise, all the preprocessing implementations can be found in the main code repository [5]. We check for duplicate images using a perceptual hash to compare images. The watermark is then removed by comparing a sample watermark image and finding the corresponding pixel location in the image.

To extract the image position, i.e., from which part of the lung an image has been taken, we make use of an optical character recognition model called Tesseract [30]. The text extraction can be found in our repository [4]. Once the positions are extracted we remove all text in the image. For that, we leverage the CRAFT model [6] which gives us bounding boxes containing text in the image. Finally,

[1] `https://pubmed.ncbi.nlm.nih.gov/`

[2] `https://www.cochranelibrary.com/`

[3] `https://www.medrxiv.org/`

[4] `https://github.com/epfl-iglobalhealth/LUS-TB-JulienVignoud`

[5] `https://github.com/epfl-iglobalhealth/LUS-COVID-main`

we crop images to remove the scale and make each image square-shaped. See appendix Figure 12a and 12b for before and after images.

4.4 Deep Learning model

DeepChest. We adapt the DeepChest model [27] because it has been tailored for lung ultrasound images using similar probes. Its original purpose is COVID-19 diagnosis and prognosis. An overview of the architecture is represented in Figure 2. The model takes as input lung ultrasound images along with the anatomic location of acquisition. Images are shaped into feature vectors using a feature extractor, in our case a pre-trained ResNet [18]. The sites embeddings are created using a positional encoding similar to the Transformers' one [31] such that the classifier can easily distinguish from which position an image comes. Both representations are added and fed into an aggregator, which combines all the images of one patient into a single representation. The original implementation uses an aggregator inspired by the natural language processing BERT model [15], which encodes relationships of words within a sentence, in their case, positions relationships for one patient. Other aggregators are Min/Max pooling, keeping the image with the most extreme feature values, or Attention Pooling, in which a neural network is trained to weigh the importance of each image of a patient. The classifier is a two-layer feed-forward neural network with $tanh$ activation. Adapting DeepChest from COVID-19 to TB diagnosis required little work as both rely on similar images as inputs.

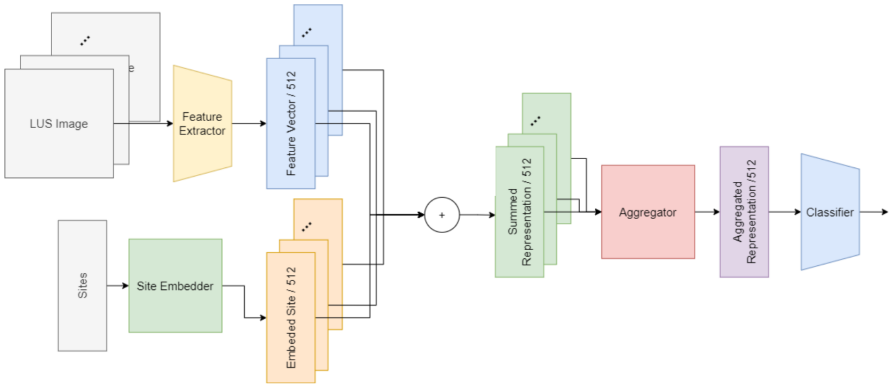


Figure 2: DeepChest model architecture. The model takes as inputs LUS images and corresponding anatomic sites. They are shaped into 512-dimensional feature vectors through the Feature Extractor and the Site Embedder respectively. The image vector and corresponding site vector are added element-wise and all the resulting representation of a single patient are combined into a single vector by the Aggregator. The classifier predicts a TB diagnosis based on a patient's representation.

Hyperparameters. The DL model hyperparameters are fine-tuned using 5-fold cross validation and evaluated on a separate test split. As one model instance is trained for each of the 5 fold, the 5 models are evaluated on the test set and will result in a mean performance along with 95% confidence interval.

Interpretability. In order to provide interpretable results, we compute class activation maps (CAM) of the last layer of the feature extractor using the Grad-CAM method [28]. By looking at the neuron activity we can measure the importance of each pixel in the diagnosis. Thus, it enables us to evaluate on which images and locations the model relies to make its predictions. The resulting CAM are assessed by LUS experts.

4.5 Baseline model from clinician-extracted features

As there are no robust estimates on the human capacity for TB diagnosis from LUS, we develop a baseline a model using handcrafted features by clinicians. Here, two double-blinded experts assess each LUS image and grade it on a an ordinal pathology scale from 0 to 6 as described in Table 3. From these values, features are engineered and explained in Table 4. The features are then fed into a

model, which will be selected and fine-tuned through cross validation among Logistic Regression, SVM and Random Forest.

We will compare the DL and the clinician models, the latter being the baseline we are aiming to reach and outperform. For a strict comparison, both models will be trained and then evaluated on the same sets, sets being different between training and testing. Furthermore, the cross-validation will be conducted for both models using the same splits.

4.6 Clinical qualitative assessment

To measure the alignment between the predictions and the physiology/pathology, LUS experts review a set of 20 predictions, one per patient. For each patient, clinicians are shown its LUS images as well the CAM overlay, similarly to Figure 11. Clinicians are given from which sites each image comes from but don't know what is the patient gold standard diagnosis nor the model prediction. Clinicians answers 4 questions for each of the 20 patients:

1. **Image Relevance:** Are the most highlighted images the most clinically relevant in the image series?
2. **Region Relevance:** Do the highlighted regions correspond to the most clinically relevant areas within each image?
3. **Physiological Alignment:** Are these regions on lung/pleural tissue?
4. What is your diagnosis?

For the first three questions, clinicians can answer with *not at all relevant*, *somewhat relevant* or *mostly relevant*, later converted to -1, 0 or 1 respectively. The scores are averaged across clinicians yielding three scores for each patient (image relevance, region relevance and physiological alignment).

The 20 patients are chosen according to the model predictions: patients are randomly selected such that there are 5 true positives, 5 true negatives, 5 false positives and 5 false negatives. By construction, the model accuracy, sensitivity and specificity on these patients are 0.5. Patients are chosen in such a way in order to analyze the CAM both when the model is correct and when it is mistaken.

4.7 Anatomic site relevance for acquisition optimisation

To reduce the number of acquired images needed, we assess the relative importance of the anatomic positions to identify the minimum number of positions (and their optimal combination) needed to reach similar performances.

We optimize the search, to avoid the costly process of iterating through all possible permutations. Firstly, we test how each position performs when used individually for predictions. Secondly, we take advantage of the attention pooling mechanism to observe the weight given to each position during validation and use it as an ordering of position usefulness. In other words, positions with higher average attention weights may be more important than those with lower values.

To exploit this ordering, we perform an experiment similar to recursive feature elimination (RFE): starting with all the positions we will eliminate the position with the least attention until only one position is left. We perform two variations of this experiment:

1. **Training RFE:** with decreasing number of positions, train and evaluate the model with the remaining positions *(Answering the question of: "What is the minimal amount of data needed for training?")*
2. **Evaluation RFE:** train the model once with all the positions and then successively evaluate it on the remaining positions without training it again. *(Answering the question of: "Given a trained model, what is the minimal amount of data needed at inference?")*

5 Results

5.1 Dataset

Each patient had an average of 24 images. According the the protocol, we anticipate 24 images per patient from 14 anatomic sites where 10 are taken in 2 standard depths, the remaining 4 lateral positions (QLD, QLG, QSLD and QSLG) are taken in 1 depth only. Figure 3 shows the frequency of each position across all patients. We can see that position representations are somewhat uneven, where QAIG, QPID and QPIG are notably fewer than anticipated, likely due to their placement close to the heart (QAIG) or difficult posterior access on supine patients unable to mobilise (QPID and QPIG).

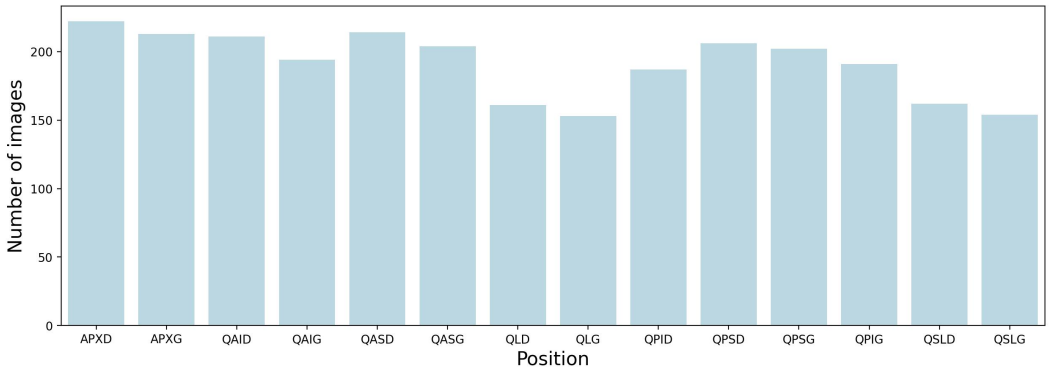


Figure 3: Distribution of anatomic position among all images

Figure 4 shows the discrepancy in the number of images per patient, ranging from 11 to 32. These disparities illustrate the need for model robustness.

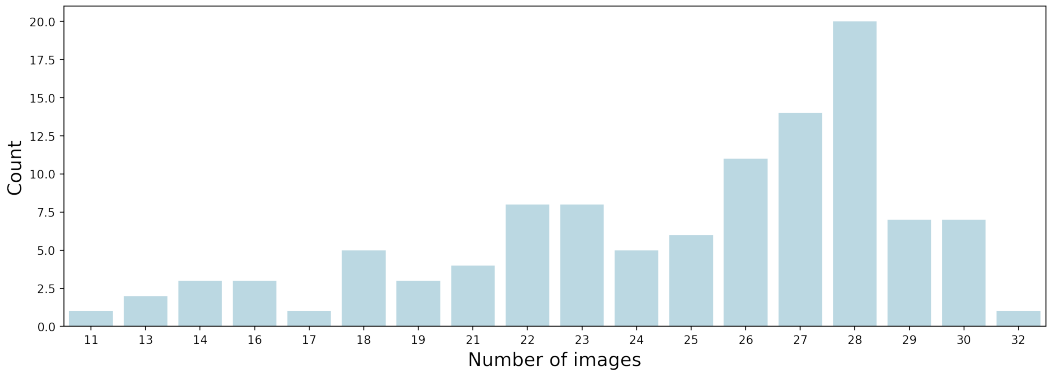


Figure 4: Distribution of images per patient

5.2 Deep learning model performance

The model was trained by optimizing the binary cross-entropy loss and evaluated with AUROC and balanced accuracy. Among the different aggregators, attention pooling yielded the best results. Another optimized hyperparameter was the independent image dropout: the proportion of input images dropped per patient. Interestingly, the best dropout was 0.7, i.e., the predictions were best when 70% of input images were dropped. While increasing the dropout, the AUROC stays stable but the 95% confidence interval across the test folds is greatly reduced. Such a result may be explained by the high correlation between same site images, multiple images being taken for each site, as well as the high variation in the number of images per patients. Indeed, a higher dropout may reduce overfitting and increase the robustness to differences between patient's images. The optimal hyperparameters found are reported in Table 5.

Furthermore, the model was evaluated on the RSA (South Africa) dataset and the different model performances are shown in Table 1 and Figure 5. The *Clinician Benin* model is discussed in the next subsection. We can see a performance difference between the two cohorts, possibly explained by overfitting, lack of generalization or differences in image acquisition. This point is further discussed in the Section 6. WHO guidelines for tuberculosis triage tests advise for 90% sensitivity and 70% specificity [25]. The specificity objective is already fulfilled by the *DL Benin* model and the sensitivity threshold is almost reached, suggesting potential for the use of such models to predict TB.

Model	**AUROC**	**Balanced Accuracy**	**Sensitivity**	**Specificity**
DL Benin	**0.88** [0.86, 0.91]	**0.81** [0.79, 0.82]	**0.84** [0.80, 0.88]	**0.76** [0.71, 0.81]
DL RSA	0.74 [0.71, 0.77]	0.69 [0.66, 0.73]	0.63 [0.53, 0.73]	0.75 [0.69, 0.81]
Clinician Benin	0.57 [0.56, 0.59]	0.51 [0.51, 0.51]	0.38 [0.38, 0.38]	0.64 [0.64, 0.64]

Table 1: Test metrics of the different models. Results are reported with the mean and 95% confidence interval.

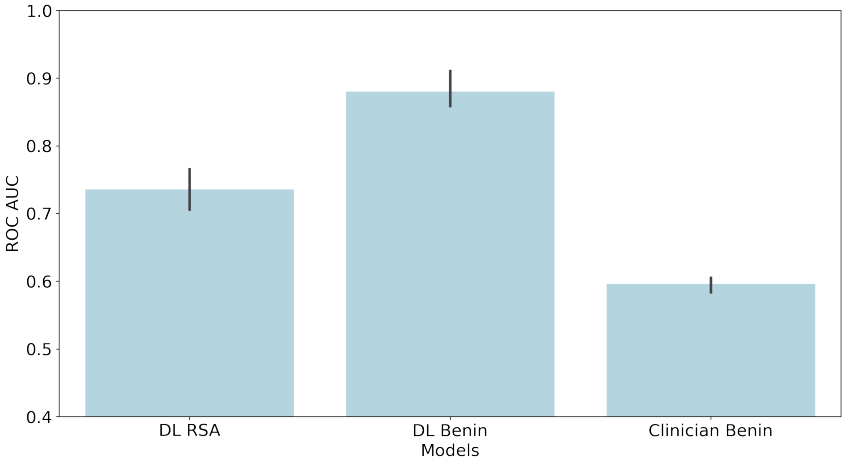
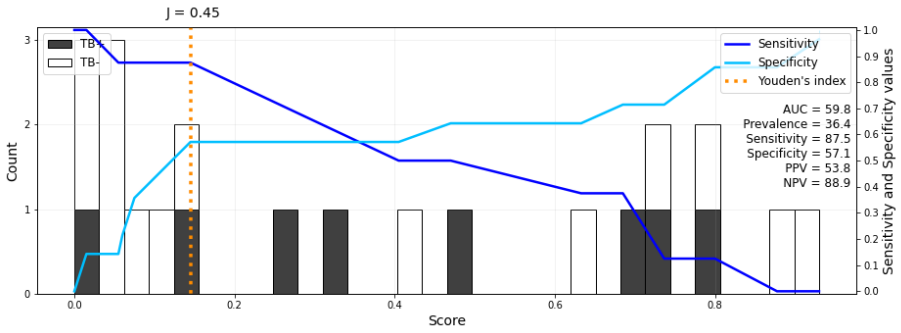


Figure 5: Model comparison based on test AUROC

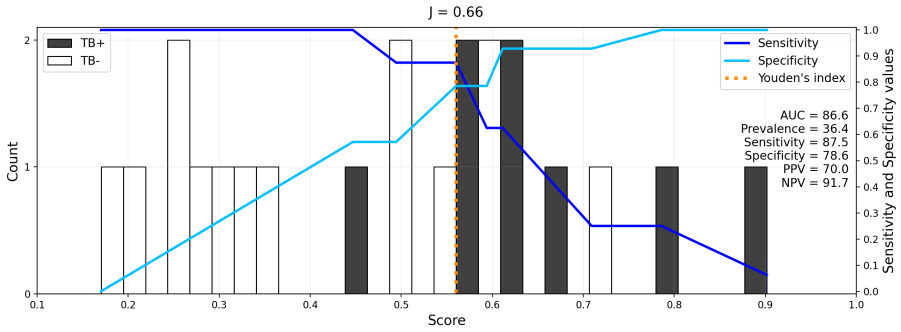
5.3 Clinician model baseline

Among random forest, logistic regression and SVM, the best model based on the clinician-extracted features was the random forest, with gini criterion, 200 trees and a maximum depth of 200. An AUROC comparison is illustrated in Figure 5 and the exact results are reported in Table 1. We can see that the result is near random with a statistically significant performance distinction between the DL model and the clinician model, the former reaching 0.9 AUROC compared to 0.6 for the latter ($p < 0.001$).

Furthermore, we can compare Youden's J statistic of both model diagnoses on Figure 6. The statistic measures the performance of a binary diagnostic test through a value between 0 and 1. A zero value denotes a useless test while a value of 1 signifies the test is perfect, without false positive nor false negative. The clinician model has a J statistic of 0.45 while the DL model diagnosis reaches 0.66. The Figure 6 shows the score predicted for each TB positive and negative patients by the two models. Patients with scores above 0.5 are predicted TB positive. We can see that more TB negative patients, denoted by white bars, are given lower scores by the DL model than by the Clinician model. We can observe a similar effect for TB positive patients.



(a) Clinician model predictions



(b) DL model predictions

Figure 6: Prediction comparisons along with main metrics and Youden's J statistic

5.4 Clinician qualitative assessment

The model predictions were reviewed by three different LUS experts, answering for each patient questions listed in Section 4.6. We only analyze positive model predictions due to the Grad-CAM mechanism: further work is needed to interpret negative predictions.

Overall, we observe in Table 2 that scores, taking value between -1 and 1, are positive. Hence, the diagnoses are considered relevant. Particularly, selected images for true positive DL predictions are systematically evaluated as relevant by clinicians. We notice similar patterns for region relevance and physiological alignment: true positives cases almost always score perfect relevance. It is interesting to note that false positives are still considered relevant despite having lower scores than the true positives locations. Hence, we can see a correlation between the diagnosis correctness and the highlighted locations relevance evaluated by clinicians.

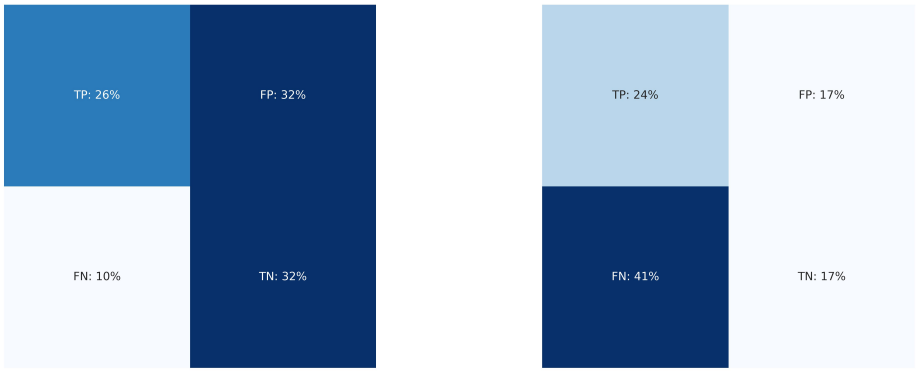
DL prediction class	Image relevance	Region relevance	Physiological alignment
TP	1	0.8	1
FP	0.53	0.13	0.53

Table 2: Overall question scores w.r.t the DL prediction class. The minimum is -1, for totally irrelevant, and the maximum is 1, for totally relevant.

Finally, the clinicians are asked to diagnose each patient. In average, they correctly diagnose 52% of the patients, compared to 50% by construction for the model. In other words, LUS experts make almost the same number of mistakes as the model on this set of patients.

However the incorrect diagnoses are not made on the same patients as the model. As we can see in Figure 7a for correct clinicians diagnoses, the model predictions are evenly distributed between true

positives, true negatives but also false positives: the model struggles more than clinicians to classify some TB- patients. For incorrect clinician answers, in Figure 7b, most patients are classified as false negatives, hence, showing that both model and clinicians are making mistakes when diagnosing some TB+ patients.



(a) For correct clinician diagnoses, the percentage of clinician answers falling in each DL prediction class (TP, TN, FP, FN)

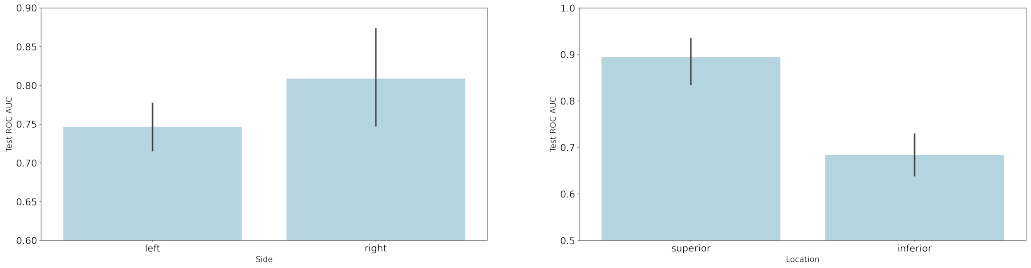
(b) Idem for incorrect clinician diagnoses

Figure 7: DL prediction class percentages when clinicians are correct vs incorrect. For instance, the percentage of the TP class for correct clinician diagnoses is the percentage of correct clinician answers predicted as TP by the model.

5.5 Site importance and combinations

As a first step, we train and evaluate each site individually. We group left and right corresponding positions as there is no medical evidence of side asymmetries in TB nor empirical ($p > 0.05$, see Figure 8a for AUROC comparison). The performances are illustrated in Figure 13. A clear site importance ordering doesn't emerge from individual explanatory power.

It is however interesting to notice the statistically significant performance difference between inferior and superior positions in Figure 8b, with $p < 0.001$. Indeed, TB is well known to preferentially occupy the upper part of the lungs [7].



(a) Left vs right sites

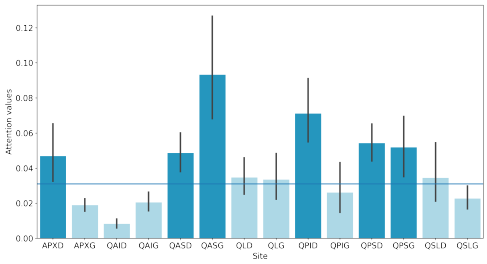
(b) Inferior vs superior sites

Figure 8: Comparison for asymmetric anatomic performances

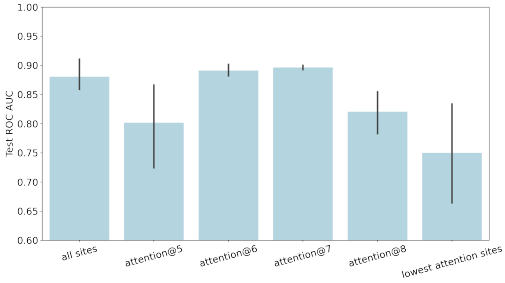
Another method to measure site importance may be achieved by looking at the attention pooling mechanism, extracting the weight attributed to each position by the model aggregator. The average attention weight attributed to each position in the validation sets is represented in Figure 9a. We can

see a clear difference in the attention given to different positions. We can notice a threshold, denoted by the horizontal line, separating the lower 95% confidence interval bound of the different positions. The 6 positions above this threshold are coloured in a darker blue. We then train and evaluate a model using only these positions. The test AUROC is reported in Figure 9b, where the 6-site model is called *attention@6*. In addition, we evaluate models with both less or more positions, following the importance ordering given by Figure 9a. Finally, we compare the results with the *all sites* model as well as the model based on the lowest attention positions for method validation.

As expected, the *lowest attention sites* model performs worse and with higher variance than the *all sites* model. While *attention@5* and *attention@8* both perform worse than *all sites* ($p < 0.001$), we observe that *attention@6* and *attention@7* matches *all sites*'s AUROC: $p > 0.1$ we cannot reject the null hypothesis. We can also note a smaller 95% confidence interval than *all sites*, maybe denoting a higher generalization power.

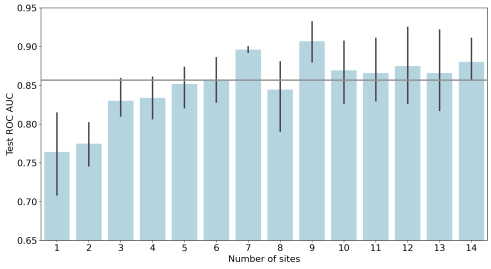


(a) Average attention weights given to each site during validation

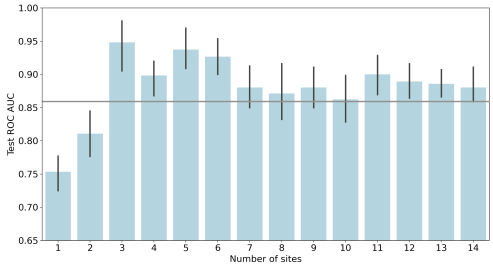


(b) Model performances w.r.t. positions used

In order to find best position combinations in a systematic manner, we proceed with *Training RFE* and *Evaluation RFE*. As a reminder, both successively eliminate the position with lowest attention weight from the input images, Training RFE trains a new model instance for each number of positions while Evaluation RFE always uses the *all sites* model. The test AUROC's for each number of positions are reported in Figure 10a and Figure 10b for Training RFE and Evaluation RFE respectively. In both cases, we observe that the lowest number of positions matching the *all sites* performance is 3. In other words, the model performs as good with 3 positions as when fed all 14 sites. This result suggests high potential for clinician data acquisition improvements.



(a) Training RFE test AUROC w.r.t. the number of positions



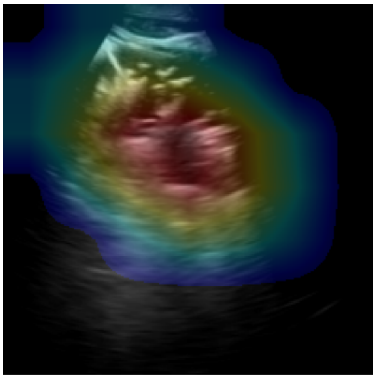
(b) Evaluation RFE test AUROC w.r.t. the number of positions for

6 Discussion

6.1 Limitations

The main limitation of this work may be the data with which models have been trained and evaluated. While a sample of a 111 patients may be sufficient to reach better performances than clinicians, it is

(a) Lung ultrasound sample image



(b) Image (a) overlaid with the class activation map, red means a high weight in the diagnosis prediction while blue or uncoloured pixels are less useful.

Figure 11: Ultrasound image side-by-side with the pixel-wise importance for the model prediction

not enough to reach the robustness necessary to overcome variability in image acquisition. To address this issue, increasing the number of patients and images is a first step and including different cohorts for training and evaluation is necessary, as shown by the performance on the Benin and RSA test sets.

The comparison between the DL model and the clinician model can be made more reliable by extracting the same features by more clinicians and aggregating results. This way, the reliability of the experiment could be measured, for instance through Cohen's Kappa, and the results less dependent of the inter-clinician variability.

Moreover, the comparison between the Benin and RSA cohorts of DL model performance still lacks a clinician baseline, which would give a human reference of the dataset diagnosis complexity. In other words, extracting clinician features on the RSA dataset would complement the performance comparison between Benin and RSA.

Finally, the gold standard of RSA may be flawed by the high HIV prevalence producing pauci-bacillary sputum samples and a high number of false negatives. More work is needed to validate these diagnoses by microbiological culture.

6.2 Future work

A line of work to consider in priority may be to improve the fine-tuning of the model hyperparameters in the objective to reach the 90% sensitivity threshold advised by WHO. As no other existing method may be compared, it is important to consider different models, for instance with less parameters or simpler architectures and assess if similar or better results can be achieved at lower computational costs.

Indeed, the Benin model trained on all sites takes 150 epochs to converge, roughly 2 hours, which is half the number of epochs required by its COVID-19 equivalent, but may still have significant ecological impact. In fact, each model training generates approximately 38.2 gCO2eq according to the CUMULATOR tool [1]. The completion of this work required around 180 such runs, where a 5-fold cross validation accounts for 5 individual runs. In total, the sum of all our experiments accounts for 6.9 kgCO2eq.

Further possibilities to improve the predictive power may reside in using ultrasound videos, with the added information of tissue displacement through time, or in tabular data such as symptoms or demographics.

Finally, regarding image acquisition variability, it has been recommended to realize the pre-training of the feature extractor on medical images rather than real-world images [23]. However, medical image databases are still multiple order of magnitude smaller than the commonly used ImageNet [13]. Another method to reduce operator variability may be to integrate real-time AI feedback during image acquisition in order to standardize the image quality.

Importantly, all patients may have co-infections/morbidities such as COVID-19, HIV or malaria. Stratified analysis may help explain the distribution of accuracy in vulnerable subgroups like HIV and diabetes.

6.3 Conclusion

Addressing the need for new TB diagnosis methods in low-resource settings as well as the lack of existing solutions in the literature, we developed a DL model meeting WHO criteria and overcoming the capacity of features handcrafted by clinicians. Moreover, the model is able to work successfully from only 3 body locations hence showing great potential to optimize the image acquisition protocol followed by clinicians. The model has been validated on a different cohort and qualitatively assessed by LUS experts. Meeting the interpretability necessity of the diagnosis, the prediction alignment with physiology and pathology can be measured by looking at location importance for the prediction. Further work is needed to explore other solutions and improve the method's robustness towards real world application.

References

[1] CUMULATOR — a tool to quantify and report the carbon footprint of machine learning computations and communication in academia and healthcare.

[2] WHO and Global Fund Warn Inequalities Block Progress Towards Ending AIDS, TB and Malaria - News - The Global Fund to Fight AIDS, Tuberculosis and Malaria.

[3] World Health Organization (WHO) COVID-19: Considerations for tuberculosis (TB) care. 2020.

[4] State of inequality: HIV, tuberculosis and malaria, 12 2021.

[5] Zeynettin Akkus, Jason Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D. Weston, Kenneth A. Philbrick, and Bradley J. Erickson. A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow. *Journal of the American College of Radiology : JACR*, 16(9 Pt B):1318–1328, 9 2019.

[6] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character Region Awareness for Text Detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:9357–9366, 4 2019.

[7] V. Balasubramanian, E. H. Wiegeshaus, B. T. Taylor, and D. W. Smith. Pathogenesis of tuberculosis: pathway to apical localization. *Tubercle and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 75(3):168–178, 1994.

[8] Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Avinash Aujayeb, Michael Moor, Bastian Rieck, and Karsten Borgwardt. Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis. *Applied Sciences 2021, Vol. 11, Page 672*, 11(2):672, 1 2021.

[9] Laura J. Brattain, Brian A. Telfer, Manish Dhyani, Joseph R. Grajo, and Anthony E. Samir. Machine learning for medical ultrasound: status, methods, and future opportunities. *Abdominal radiology (New York)*, 43(4):786–799, 4 2018.

[10] Danielle Cazabon, Tripti Pande, Sandra Kik, Wayne Van Gemert, Hojoon Sohn, Claudia Denkinger, Zhi Zhen Qin, Brenda Waning, and Madhukar Pai. Market penetration of Xpert MTB/RIF in high tuberculosis burden countries: A trend analysis from 2014-2016. *Gates Open Research*, 2, 2018.

[11] Centers for Disease Control and Prevention. Chapter 4 Diagnosis of Tuberculosis Disease Chapter Objectives.

[12] Payal Dande and Purva Samant. Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review. *Tuberculosis*, 108:1–9, 1 2018.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. pages 248–255, 3 2010.

[14] Prabha Desikan. Sputum smear microscopy in tuberculosis: Is it still relevant? *The Indian Journal of Medical Research*, 137(3):442, 3 2013.

[15] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018.

[16] Sanam Ebrahimzadeh, Nayaar Islam, Haben Dawit, Jean-Paul Salameh, Sakib Kazi, Nicholas Fabiano, Lee Treanor, Marissa Absi, Faraz Ahmad, Paul Rooprai, Ahmed Al Khalil, Kelly Harper, Neil Kamra, Mariska MG Leeflang, Lotty Hooft, Christian B van der Pol, Ross Prager, Samanjit S Hare, Carole Dennie, René Spijker, Jonathan J Deeks, Jacqueline Dinnes, Kevin Jenniskens, Daniël A Korevaar, Jérémie F Cohen, Ann Van den Bruel, Yemisi Takwoingi, Janneke van de Wijgert, Junfeng Wang, Elena Pena, Sandra Sabongui, Matthew DF McInnes, and Cochrane COVID-19 Diagnostic Test Accuracy Group. Thoracic imaging tests for the diagnosis of COVID-19. *Cochrane Database of Systematic Reviews*, 2022(5), 5 2022.

[17] Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. Machine Learning for Medical Imaging. *Radiographics : a review publication of the Radiological Society of North America, Inc*, 37(2):505–515, 3 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:770–778, 12 2016.

[19] Zhaoyu Hu, Zhenhua Liu, Yijie Dong, Jianjian Liu, Bin Huang, Aihua Liu, Jingjing Huang, Xujuan Pu, Xia Shi, Jinhua Yu, Yang Xiao, Hui Zhang, and Jianqiao Zhou. Evaluation of lung involvement in COVID-19 pneumonia based on ultrasound images. *Biomedical engineering online*, 20(1), 12 2021.

[20] Sangheum Hwang, Hyo-Eun Kim, Jihoon Jeong, and Hee-Jin Kim. A novel approach for tuberculosis screening based on deep convolutional neural networks. *Medical Imaging 2016: Computer-Aided Diagnosis*, 9785:97852W, 3 2016.

[21] Sagar Kulkarni and Saurabh Jha. Artificial Intelligence, Radiology, and Tuberculosis: A Review. *Academic Radiology*, 27(1):71–75, 1 2020.

[22] Paras Lakhani and Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2):574–582, 8 2017.

[23] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering*, 5(2):261–275, 4 2019.

[24] Sizulu Moyo, Farzana Ismail, Martie Van der Walt, Nazir Ismail, Nkateko Mkhondo, Sicelo Dlamini, Thuli Mthiyane, Jeremiah Chikovore, Olanrewaju Oladimeji, David Mametja, Phaleng Maribe, Ishen Seocharan, Phumlani Ximiya, Irwin Law, Marina Tadolini, Khangelani Zuma, Samuel Manda, Charalambos Sismanidis, Yogan Pillay, and Lindiwe Mvusi. Prevalence of bacteriologically confirmed pulmonary tuberculosis in South Africa, 2017–19: a multistage, cluster-based, cross-sectional survey. *The Lancet Infectious Diseases*, 0(0):2017–2036, 5 2022.

[25] Ruvandhi R. Nathavitharana, Christina Yoon, Peter Macpherson, David W. Dowdy, Adithya Cattamanchi, Akos Somoskovi, Tobias Broger, Tom H.M. Ottenhoff, Nimalan Arinaminpathy, Knut Lonnroth, Klaus Reither, Frank Cobelens, Christopher Gilpin, Claudia M. Denkinger, and Samuel G. Schumacher. Guidance for Studies Evaluating the Accuracy of Tuberculosis Triage Tests. *The Journal of infectious diseases*, 220(220 Suppl 3):S116–S125, 10 2019.

[26] L. J. Salomon, N. Winer, J. P. Bernard, and Y. Ville. A score-based method for quality control of fetal images at routine second-trimester ultrasound examination. *Prenatal Diagnosis*, 28(9):822–827, 9 2008.

[27] Hugo Schmutz. DeepChest: A neural attention model for interpretable, missingness-resilient diagnosis and risk stratification of COVID-19 from lung ultrasound images. *Unpublished*, 2020.

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:618–626, 12 2017.

[29] Hamid Shokoohi, Maxine A. Lesaux, Yusuf H. Roohani, Andrew Liteplo, Calvin Huang, and Michael Blaivas. Enhanced Point-of-Care Ultrasound Applications by Integrating Automated Feature-Learning Systems Using Deep Learning. *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine*, 38(7):1887–1897, 7 2019.

[30] Ray Smith. An overview of the tesseract OCR engine. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2:629–633, 2007.

[31] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need.

[32] L. Wang, M. O. Turner, R. K. Elwood, M. Schulzer, and J. M. FitzGerald. A meta-analysis of the effect of Bacille Calmette Guérin vaccination on tuberculin skin test measurements. *Thorax*, 57(9):804–809, 9 2002.

[33] Juerd Wijntjes and Nens van Alfen. Muscle ultrasound: Present state and future opportunities. *Muscle and Nerve*, 63(4):455–466, 4 2021.

[34] World Health Organization. Global tuberculosis report. *Global tuberculosis report*, 2021.

[35] Lingyi Zhao and Muyinatu A. Lediju Bell. A Review of Deep Learning Applications in Lung Ultrasound Imaging of COVID-19 Patients. *BME Frontiers*, 2022:1–17, 2 2022.

A Appendix

Code	Interpretation
0	A-lines, normal lung sliding
1	B-lines (3 or more per field)
2	Coalescent B-lines
3	Subplerual pathology smaller than 1cm
4	Consolidation greater than 1cm
5	A-lines, absent lung sliding (pneumothorax suspicion)
6	Pleural effusion

Table 3: Code meaning attributed to each site by clinicians during image acquisition

Type	Features	Interpretation
Binary Features	Any quadrant with 0	1 if the patient has an image with value 0, 0 otherwise
	Any quadrant with 1	
	Any quadrant with 2	
	Any quadrant with 3	
	Any quadrant with 4	
	Any quadrant with 5	
	Any quadrant with 6	
	Any quadrant with >=1	1 if the patient has an image with value greater or equal to 1, 0 otherwise
	Any quadrant with >=2	
	Any quadrant with >=3	
	Any quadrant with >=4	
	Any quadrant with >=5	
	Bilateral >1	1 if the patient has one location where both the left and right positions have non-zero value
	Unilateral >1	1 if the patient has one position with non-zero value
	Apical >1	1 if the patient has one apical position (upper part of the lung) with non-zero value
Continuous features	Number of quadrants with 0	
	Number of quadrants with >=1	
	Number of quadrants with >=2	
	Number of quadrants with >=3	
	Number of quadrants with >=4	
	Number of quadrants with >=5	
	Number of quadrants with >=6	
	Total score	Sum of all the image values attributed to the patient

Table 4: Baseline model's handcrafted features

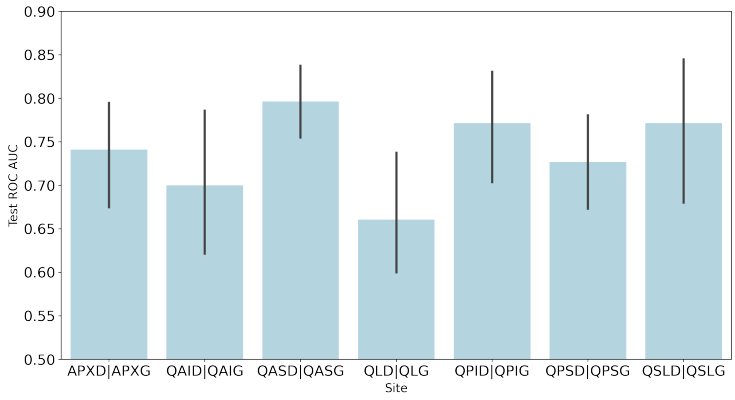
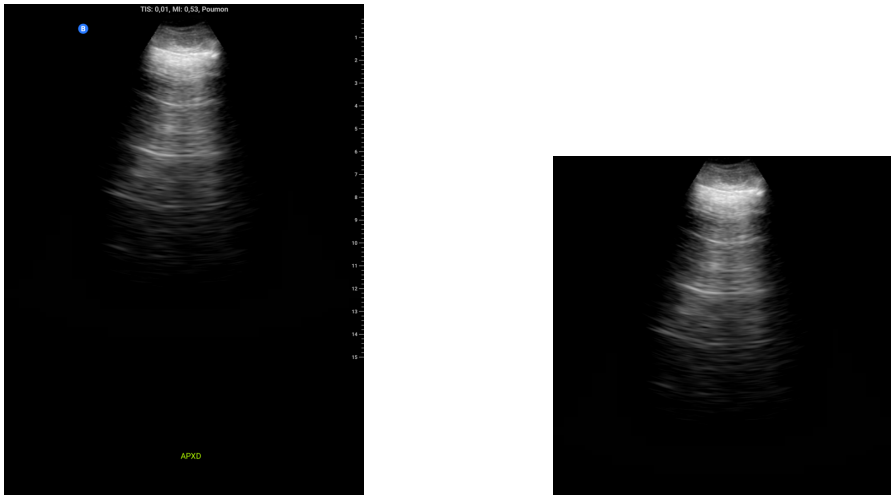


Figure 13: Individual site performances



(a) Lung ultrasound original image

(b) Image (a) after preprocessing

Figure 12: Ultrasound before and after preprocessing

Parameter	**Fine-tuned value**
Batch size	32
Learning rate	0.0001
Aggregation type	`MLP_AttentionPooling`
Epochs	150
Independent dropout	0.7
Weight decay	0.0
Minimized metric	Validation BCE loss
Embedding dimension	512

Table 5: Result of the hyperparameter fine-tuning