VINBRAIN

FREUNDESKREIS FÜR INTERNATIONALE TUBERKULOSEHILFE

Utilizing Artificial Intelligence in reading Chest X-rays for Tuberculosis screening in Vietnam

“Tuberculosis is the perfect expression of our imperfect civilization.”

Table of Contents

1. Introduction

Despite the existence of rapid diagnostic tools, highly effective and cost-effective treatment regimens and viable means to prevent the disease, TB still kills 1.4 million people per year. This makes TB second only to COVID-19 as a cause of deaths worldwide from an infectious disease. Key facts:

1. Just 30 high TB burden countries account for an estimated 87% of the world's TB burden. Eight countries account for two-thirds of the total, with India leading the count, followed by Indonesia, China, the Philippines, Pakistan, Nigeria, Bangladesh and South Africa. Viet Nam ranks 11th in the list of high TB burden countries;
2. Between 2015 and 2019, the global TB incidence rate declined by just 9%, less than half of the reduction target in the End TB Strategy;
3. Ending the TB epidemic by 2030 is among the prioritized health targets of the United Nations Sustainable Development Goals (SDGs).

Vietnam's success in handling COVID-19 pandemic has proved that the country has the determination and resilience to overcome the most burdensome public health epidemics. However, "in order to achieve the objectives of this plan [to end TB], besides strengthening current interventions, Vietnam must harness new technologies and innovation." - Dr. Kidong Park, Representative WHO Vietnam.

One of the pathfinding developments in TB is the application of AI to enhance efforts to find and treat more persons with TB, and thereby cut down community transmission to a level at which the disease no longer represents a public health burden.

This white paper aims to provide the reader with a better understanding of:

1. The current situation of TB diagnosis in Vietnam;
2. VinBrain's AI solution and its development process; and
3. Next steps in supporting Vietnam's progress towards ending TB.

2. Tuberculosis in Vietnam

Vietnam ranks 11th among the world's 30 highest TB burden countries. In 2018, the country notified 174,000 drug-sensitive TB patients and reported 13,200 TB-related mortalities.[1] There were an estimated 8,600 people with drug-resistant TB, but just 3,110 (36%) were started on treatment. In 2014, the Government of Vietnam passed legislation codifying the country's strategy to end TB.[2] However, there are key gaps in the country's current TB response, which are preventing this ambitious goal from being achieved.

Figure 1: Health workers from the National Lung Hospital distribute free medicines to disadvantaged locals in rural and remote areas in Thanh Hoa province

The second national TB prevalence survey (conducted in 2017/18) showed that nearly half of Vietnamese TB patients remain unreached by the government's National TB Control Program (NTP).[3] A recent inventory study indicated that these missed individuals are roughly evenly split into those who are undiagnosed (truly missed) and those who are treated, but never reported to the NTP.[4] The individuals who are truly missed often have limited access to healthcare services, due to physical barriers (e.g. distance from home to the health facility), cost (out of pocket spending and income loss), convenience, perceived quality, and stigma, which are associated with utilizing the government's TB services.

There is an increase in the number of TB patients using private-sector healthcare services in Viet Nam as the country's economy grows. Yet, under 5% of notifications originate from the private sector. The quality of private-sector TB care is highly variable, which could be contributing to the TB epidemic as people fail to complete treatment successfully or relapse after inadequate treatment. Meanwhile, a recent survey of TB-affected households showed that two-thirds incurred catastrophic costs (income loss >20% of annual household income).[5] These data convincingly show that TB care in Vietnam is not yet patient-centered and that this disease is a key driver of poverty.

The annual rate of TB infection is 1.7% and an estimated 30% of Vietnamese population has a latent TB infection (LTBI).[6] An estimated 5-10% of people with LTBI may develop active TB disease in their lifetimes, serving as a vast reserve for future TB disease, even if the new transmission were eliminated.

As a result, the NTP has identified the need to intensify innovations to find, treat and support more persons with TB and LTBI to end TB by 2030.[7] A key component of these innovations is the optimal application of new tools and technologies.

3. The current diagnostic paradigm

The Nation TB Control Program estimates that 62% of people with active, transmissible TB disease would have been missed if they had been screened using the standard diagnostic algorithm, which relied on verbal screen for prolonged cough. Furthermore, surveys have shown that a substantial proportion of individuals with TB are entirely asymptomatic and only indicated for testing through chest X-ray (CXR) screening.[8] The large number of missed individuals with TB contributes to sustained community TB transmission.

A major barrier to treating more people with TB is the traditional reliance on smear microscopy for bacteriologic confirmation and the need for smear-negative individuals to undergo a lengthy clinical evaluation process with exploratory antibiotic treatment, follow-up visits to the clinic and additional CXR and sputum tests before treatment is initiated.

To accelerate the reduction of TB, the country is transitioning to a new molecular diagnostic tool called Xpert MTB/RIF assay (Xpert). Xpert was heralded in 2010 as a 'game-changer' for TB diagnosis.[9] This test has a higher sensitivity than smear microscopy, and is able to simultaneously detect resistance to rifampicin, which is a surrogate marker for multidrug-resistant TB (MDR-TB). Following endorsement by the World Health Organization (WHO) and a price reduction for high-burden and low-income countries, the test has been rapidly scaled up.[10]

Figure 2: Doctor executes TB diagnostic test using GeneXpert MTB/RIF

Early hopes that Xpert scale-up would be sufficient to significantly increase case detection under passive case finding models have been dampened by reports from several countries, such as South Africa[11] and Nepal.[12] These evinced that wide-scale Xpert testing does not increase case detection due to the large numbers of patients treated empirically and based on clinical evidence in the traditional diagnostic algorithm. It is now clear that an expansion in the numbers of people tested for TB, in conjunction with the shift to Xpert, is needed in order to improve detection and treatment numbers.

4. Chest radiography in tuberculosis detection

4.1. International guidance and national strategy

WHO recently released updated guidelines[13] identifying the roles that CXR can play in the detection of pulmonary TB, including as a triage tool, a diagnostic aid or an initial screening test.

Vietnam's National TB Control Program (NTP) has developed a comprehensive National Strategic Plan for 2021-2025, which includes a broad-scale transition from smear microscopy to Xpert as the first-line diagnostic tool for the detection of TB. To mitigate the resource requirements of this transition, the NTP is scaling up its "Double-X" diagnostic algorithm (X-ray screening/Xpert testing) across all 63 provinces. This algorithm indicates Xpert testing for all persons with parenchymal abnormalities on chest radiography. However, factors such as high variability in human scoring, poor specificity and limited number of appropriately trained radiographers may limit the potential utility of CXR in some settings of Vietnam, particularly in rural and remote areas.

Modelling of different diagnostic algorithms by WHO further showed that in the context of transitioning to the Xpert diagnostic test, integration of CXR as a screening step to replace the clinical symptom screening can generate both a substantially higher yield, while keeping cost escalation at a moderate, acceptable level.

Table 1: Different diagnostic algorithms and actual total yield

	CXR Screens	AFB Tests	Xpert Tests	Estimated TB Yield	Marginal TB Yield	Total Diagnostic Costs	Cost per Bac(+) Case Detected
Cough ≥2 weeks followed by smear microscopy	0	9430	0	65	-	9915	153.12
Any cough followed by smear microscopy	0	15,101	0	91	+40.6%	10,941	120.16
Cough ≥ 2 weeks followed by Xpert	0	0	9430	128	+97.7%	113,070	888.36
CXR abnormal followed by smear microscopy	34,529	4722	0	130	+100.0%	57,620	444.92
Any cough followed by Xpert	0	0	15.101	180	+178.0%	176,131	978.51
CXR abnormal followed by Xpert	34,529	0	4722	256 *	+295.3%	109,274	426.85

Source:

https://pubmed.ncbi.nlm.nih.gov/33321696/

https://www.nature.com/articles/s41598-019-51503-3.pdf

4.2 Artificial Intelligence and computer-aided reading for TB

Artificial Intelligence, also known as AI, is one of the most rapidly growing technologies today, and "it is going to change the world more than anything in the history of mankind" – according to AI oracle and venture capitalist, Dr. Kai-Fu Lee, 2018. It enables computers to discover and learn from data without being programmed. An important goal of AI is to make machines that can take data inputs, make decisions and take action as part of a loop of perception, cognition and learning. Especially, in the healthcare or medical field, AI has a tremendous contribution in reducing repetitive tasks and assisting doctors to quickly and accurately make final decisions.

The application of AI and deep neural networks has also proliferated in the health sector. One of the early use cases

consisted of the emergence of computer-aided detection (CAD) software which leveraged machine algorithms to enhance the accuracy of CXR interpretation. Specifically, these software products were trained to analyze digital CXR images for abnormalities to quantify the likelihood of active TB disease in the patient. While these technologies remain – at least politically – far from replacing human doctors, they can serve as unbiased diagnostic support tools for early-career radiologists and seasoned experts alike, when screening for pulmonary TB. These tools can further ameliorate the risk of inter-reader variability and reduce delays in the reading and interpretation of CXR in the absence of trained personnel.

4.3 Key considerations in employing CAD for TB

A key advantage of these tools is that their results are usually expressed as a firm, quantitative score sometimes raging either from 1 to 100 or from 0.01 to 0.99. These scores are subsequently measured against a threshold to inform further diagnostic evaluations. The challenge is that these thresholds themselves are complex constructs and their predictive diagnostic accuracy often depends on numerous external factors. They may even depend on the type of radiology equipment used and the clarity of image produced, vary across different CAD products and may even show substantial variability within the same software as it matures and continues to learn from new data-points.

As such, threshold values are dynamic and their use ideally requires prior calibration for the local context and differentiated definition across varying patient populations. The absence of such differentiated approach would have considerable implications at best on the reliability of the CAD and at worst on the efficiency and performance of national TB control programs. For example, thresholds that enable a program to detect TB patients with greater sensitivity will also likely incur higher diagnostic costs. Conversely, thresholds set to be more accurate in detecting true negative cases will save diagnostic costs for follow-up testing, but will also likely miss more TB cases (Table 2). As such, conducting an accurate calibration of the software to determine locally appropriate threshold scores is therefore almost as integral for the advancement of these tools as the development of an accurate tool itself.

Table 2: Sensitivity vs Specificity

Sensitivity	Ability to correctly identify true positive cases (rule out)	100% sensitivity – 100% of true positives will be identified	Highly sensitive tests may misclassify some true negatives and require unnecessary follow up testing
Specificity	Ability to correctly identify true negative cases (rule in)	100% specificity – 100% of true negatives will be identified	Highly specific tests may misclassify some true positives resulting in missed cases and under-diagnosis

5. AI model development life cycle

VinBrain has decided to dedicate substantial resources to the fight against TB through the development of a local CAD solution for AI-supported CXR reading. As part of this effort, VinBrain had to design its own AI model development life cycle, which is outlined below.

To keep up with the continuous changes and seize the opportunities inherent in the AI revolution, all AI companies need to build a fast-paced AI life cycle. The VinBrain team has built its own AI life cycle which consists of several steps. The first step is the design phase, in which the business goal is translated into design requirements for software engineers, DevOps and IT operations engineers. This step is detailed in Section 5.1. The parallel step is to define how to collect and label data used to build the AI model. This step is described in Section 5.2. The next step delineated in Section 5.3 explains how the labeled data must be passed through pre-processing step - one of the most important steps that helps to improve the model quality. The processed data will then be transformed and fed into selected AI models. Section 5.4 is the discussion on model evaluation and how to design suitable metrics to assess model quality.

Data-driven projects only create real business (and public health) values when their AI models and their applications are deployed in production instead of sitting on the shelf. Moreover, monitoring a deployed model and its feedback loop is also a crucial step that helps to ensure the consistency and robustness of the AI system. Without this step, AI systems may lose the trust of the end-users.

5.1 System design

VinBrain has built DrAid™'s system using Cloud Computing technologies, also known as internet-based computing, which provides information and shared resources to the computer and other devices on-demand. Cloud Computing also provides on-demand access to a shared pool of configurable computing resources. It provides enterprise users with various capabilities to store and process their data in third-party data centers. Its primary focus is to maximize the effectiveness of shared resources. Cloud Computing helps companies avoid upfront infrastructure costs, and thus focus on projects that differentiate their businesses instead of infrastructure. Cloud Computing now becomes a high-demand service or utility thanks to the advantages of high computing power, low cost of services, high performance, scalability, accessibility, and availability. Many companies now have their preferred cloud service providers.

Microsoft Windows Azure is a cloud computing platform released on 1st February 2010 by Microsoft. It helps build, deploy and manage applications and services through a global network of Microsoft-managed datacenters. Providing both PaaS and IaaS services, it supports many different programming language tools and frameworks, including Microsoft-specific and third-party software and systems.

The main drivers to migrate the business to the cloud include the increased stability over traditional on-site servers and mitigation of data loss risk. Another driver is scalability: a rapidly expanding model is a big challenge that requires the best technology and platform.

There are several key solutions in DrAid™'s system design, such as:

- Improve development process by automating the build-and-release processes and centralize it in one location rather than letting it done by individual developers from their workstations;
- Use Azure Git repositories to create new features for the company;
- Continuously build and test the source code from the repository;
- Utilize Static website hosting in Blob Storage for serving static content like HTML, CSS, JavaScript, and image files directly from a storage container;
- Accelerate load times, cut latency, and improve the user experience for dynamic web applications and websites by using Azure CDN;
- Deploy one or more Docker containers to package new releases of a software product and store them in Azure Container Registry;
- Simplify the process of deploying a Kubernetes cluster by using ASK Cluster;
- To enable valuable insights and drive immediate actions in creating executive dashboards; Power BI will accelerate this process.

DrAid™'s system architecture design is described in Figure 3.

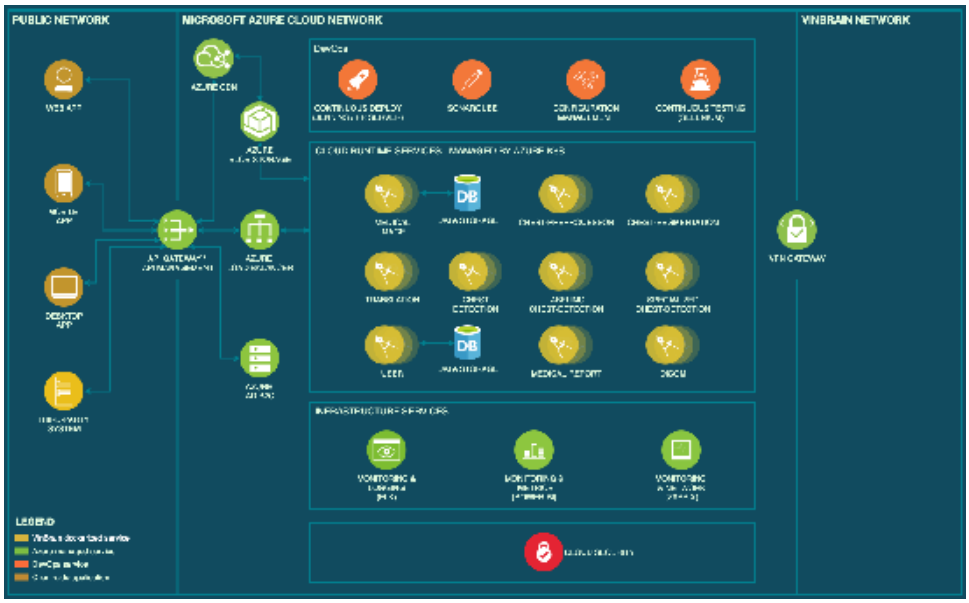

Figure 3: DrAid™ - High-level System Design

5.2 Data collection and labeling

In the field of AI, large amounts of data are required to train and fine-tune the various machine learning models. Hence, the data need to be collected and labeled.

The process of collecting and labeling data is the top priority in training machine learning models, as it enables machine learning models to gain an accurate understanding of real-world conditions. The quantity and quality of data used to train the AI will also be the major differentiation factor in terms of performance across competing platforms. First, raw collected data need to be cleaned during the data cleaning process: unqualified, duplicated and irrelevant images are excluded from the machine learning model. Data cleaning is a crucial step to improve data quality and requires a long processing time; the absence of this step impairs the performance of the machine learning model. Cleaned data then are labeled into meaningful data providing the machine learning model with real-world conditions to learn from.

There are two basic steps in data labeling. The first step is classification, where CXR images are labeled as TB positive or TB negative, then used for training binary TB classification model. The second step is segmentation, where the borders of the damaged lung areas caused by TB are collected to train the TB segmentation model.

Labeled and segmented datasets help train machine learning models to identify and understand the recurring patterns in the input to deliver accurate output. After being trained by annotated data, machine learning models can recognize the same patterns in never-before-seen data sets. As the quality and quantity of training data directly contribute to the success of an AI algorithm, developing a high-quality training dataset can be a very resource-intensive process. The following sections summarize the three common basic steps of this process:

1. Data preparation: data collection, data cleaning and radiologist expert recruitment;
2. Data labeling: classification and segmentation; and
3. Labeling evaluation.

5.2.1 Data preparation

In total, more than 500,000 medical images from many different locations in Vietnam, India and China were collected. Our team then cleaned up raw medical imaging data collected from hospitals by removing substandard images. In detail, unqualified images include images with resolution lower than 112*112 pixels, duplicated images, CT scans, MRIs, CXR of children under 16 years old, or non-anterior CXR images.

Figure 4: Unqualified X-ray image – broken image

Figure 5: Lateral CXR image

Figure 6: CXR of child under 16 years old

After filtering out the poor-quality images, we have 185,486 clean images including 11,735 positive images and 173,751 negative images. The retrospective method is the key for rapid data collection from participating lung hospitals in Vietnam and well-known public datasets (Table 3). The number of collected images shows the distribution of TB positive and negative images from each source.

Table 3: Number of clean CXR images that have been collected

Source	#Positive	#Negative
Lung Hospital 1	1,820	521
Lung Hospital 2	1,219	1,309
Lung Hospital 3	830	1,680
Lung Hospital 4	658	580
Lung Hospital 5	115	295
Lung Hospital 6	2,396	14,378
Lung Hospital 7	188	380
Lung Hospital 8	1,168	0
Lung Hospital 9	562	1,258
Lung Hospital 10	0	4,519
General clinic 1	0	18,536
Private Hospital 1	0	36,492
Private Hospital 2	134	85,797
Public dataset 1	840	7,600
Public dataset 2	359	406
Public dataset 3	1,446	0
Total	**11,735**	**173,751**

CXR image data came from various sources, including hospitals with different X-ray machine quality and imaging protocols; thus, there was substantial heterogeneity in the properties and quality of the CXR images. Hence, image pre-processing is a crucial step to acquire the best machine learning quality. In fact, the CXR images from mobile

X-ray vehicles are significantly different from the CXR images taken by X-ray machines in hospitals.

The team has analyzed and added more CXR images from both X-ray vehicles and X-ray machines to the dataset to make it more diverse. By combining collected data from hospitals and CXR scanners, the diversity and generalization of data is improved.

In the data collection process, in compliance with data protection rules and principles, all patients' information such as name, date of birth, and gender that are parts of the X-ray are removed when CXR images are fed into the model and database. De-identification is a two-step process in which our labeling tool automatically removes such information from the CXR in the first step. Then, in the second step, the radiologists review the images again; thus the CXR is completely anonymized before being fed into the system.

The radiologists who work in central and provincial lung hospitals with an average of 20 years of experience are recruited into our radiologists' team. These experienced radiologists' sensitivity and specificity are evaluated by the ground-truth test set of retrospective CXR image data (Figure 7).

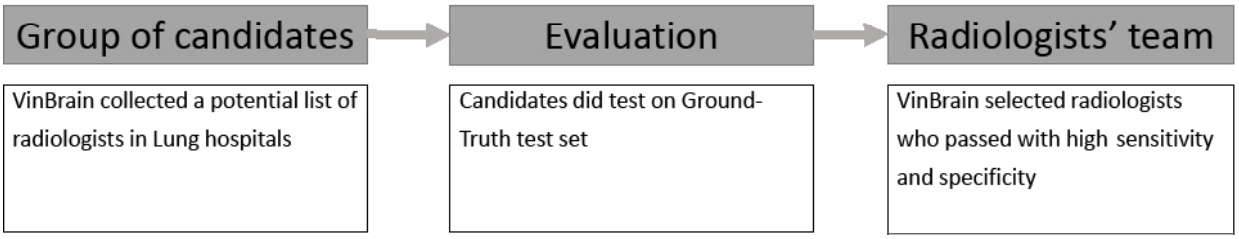

Figure 7: Process for recruiting labeling radiologists

In fact, to ensure the consistency rate between the first and second labeling of a certain radiologist, we randomly checked 10%-30% of the labeled images. The consistency rate is measured by F1 Score and Intersection over Union[14] of the labeled dataset, with the first labeled dataset assumed to be the ground truth. The labeling is expected to be unaffected by external factors and the radiologist's physical and mental conditions. For a labeled dataset to be considered as sufficient quality, it must have a consistency rate of >80%. The result is shown below:

Performance	First labeled dataset is assumed to be the ground truth The consistency is measured by F1 Score and IoU of 10% of labeled dataset
Results	F1 Score: 96% IoU: 83%

Figure 8: Evaluation of data labeling process

5.2.2 Data labeling

The critical challenges of the CXR image segmentation are the unavailability of a large number of annotated images, the ambiguity between abnormal findings, high variance of position/shape/size of a single abnormal finding, and the overlay of different lesions in the same areas. The team has solved these challenges by labeling a large dataset with several technical solutions.

Data labeling plays a significant role in the supervised learning model. The quality of data labeling directly affects the quality of model. Therefore, our labeling tool (Figure 9, Figure 10) has been created to ensure the quality of data labeling. This tool provides a CXR image viewer and helps the expert team quickly conclude and localize the abnormal area(s) with high accuracy.

Figure 9: Two doctors using VinBrain's labeling tool agreed on abnormal borders

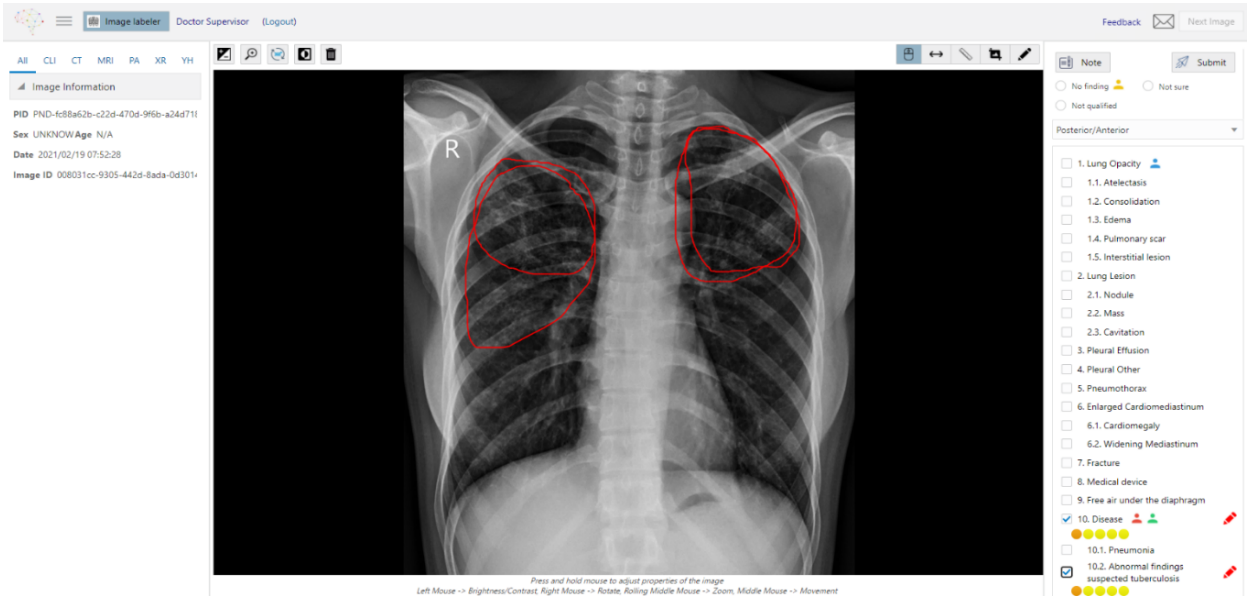

Figure 10: Two doctors using VinBrain's labeling tool disagreed on abnormal borders.

5.2.3 Labeling evaluation

After the data labeling process, the quality of labeled data is evaluated. The labeling evaluation helps reduce noise caused by human errors and makes data clearer. In this step, only high-quality labels getting a high consensus of two independent doctors and satisfy two criteria of selecting high-quality labels for machine learning are selected (Table 4). The main selection criteria include the Consistency rate and IoU.

Table 4: Criteria for evaluating label's quality

TB model	Activities of labeling	Requirements for training data	Golden set from the labeled batch
Segmentation	Border	IoU > 0.6	Confirmed by ground-truth positive info (provided by MOH)[15] Consensus on IoU (>0.6).

5.3. Pre-processing and model training

The labeled data must go through the data analysis process before being used for AI model training. This process involves pre-processing of the input images, such as data pre-processing, data transformation, and data integration. As mentioned previously, data integration combines heterogeneous data from different sources into a single scheme, providing users with a unified view of them. After this process, all cleaned/pre-processed data will be fed into selected AI models for the training process.

At the pre-processing step, a lung segmentation model based on U-net[16] network and trained with lung area annotation data, is used to crop lung area images, thus eliminating confounding factors or unrelated parts such as bone or soft tissue (Figure 10). Then, all cropped lung area images are resized at the data transformation step and used as input for training the model. By focusing only on the lung region, the performance of the TB model can be significantly improved.

In practice, all pre-processed data are split into three subsets to develop the TB classification model. 185,486 clean images were then used for the process of model training – validation – testing, and split into a training set (9,108 positive images and 161,814 negative images), validation set (1,086 positive images and 4,041 negative images), and testing set (1,541 positive images and 7,896 negative images) (Table 5).

Table 5: Number of CXR images that have been used in the development of Vinbrain's AI model for TB

Dataset	Training			Validation			Testing		
	Positive	Negative	Sum	Positive	Negative	Sum	Positive	Negative	Sum
Number of CXR images	9,108	161,814	170,922	1,086	4,041	5,127	1,541	7,896	9,437

For the TB segmentation task, the team uses k-fold cross-validation[17] of annotated data of 1,396 positive images and 20,000 negative images to build and evaluate the model. The goal of cross-validation is to test the model's ability to predict new data that was not used for the initial training, in order to flag problems like overfitting or selection bias[18] and to give an insight of how the model may perform with an independent dataset (e.g., an unknown dataset, for instance from a real problem).

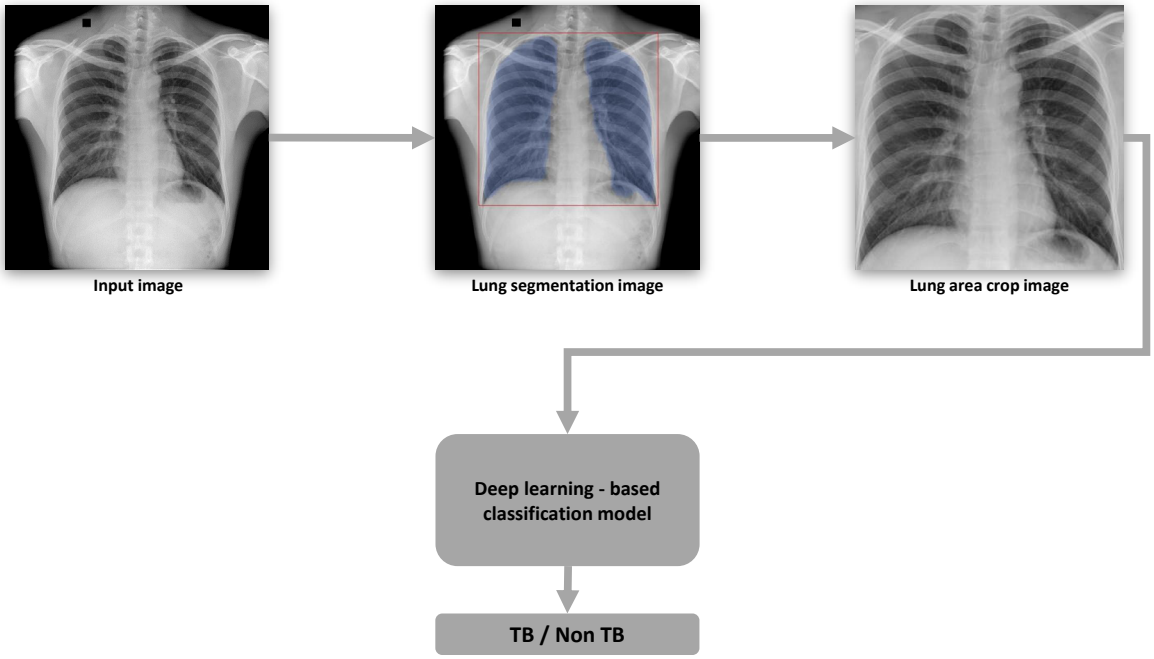

Figure 11: Tuberculosis classification approach

For TB classification, a deep learning model is created and trained using the Knowledge Distillation (KD) method with the utilized result of an ensemble model of several state-of-art Convolutional Neural Networks (CNNs). The ensemble model is a machine learning technique that combines several base models to produce one optimal predictive model. Different single pre-trained CNN models such as Densenet121, Densenet169, Densenet201[19], Xception[20], ResNext-101[21], Efficient-Net B3, Efficient-Net B5[22] have been experimented on our tuberculosis data. The top seven models are then selected for the ensemble model to improve performance. The performance improvement of the ensemble method is visible, although more difficult and expensive. To utilize the result of the ensemble of the model, the VinBrain team has applied a technique called the Distillation Method to optimize latency, model parameter and memory footprint of the ensemble model. This uses a single model Efficient-net B5 to replace the ensemble with an expected and acceptable drop in prediction accuracy (experiment results are shown in Table 5). The knowledge distillation is a promising subclass of model compression, which trains a fewer-parameters model (student) based on distillated information from the output of a larger model (teacher) on the same task to reduce the size of the model while retaining accuracy. The training process of the student model is described in Figure 12.

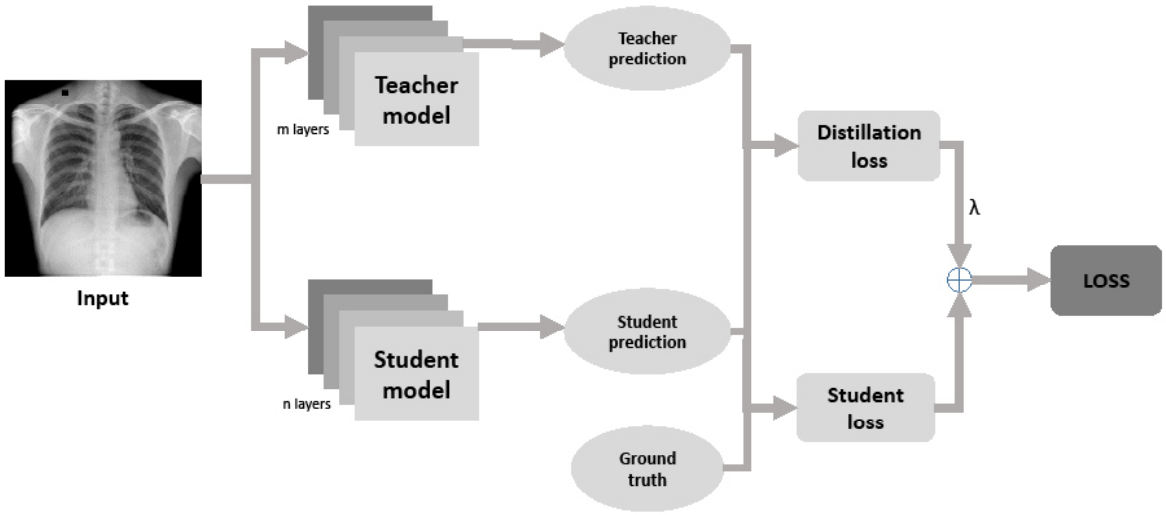

Figure 12. Knowledge distillation-based training process

In the knowledge distillation method, the student loss is the binary cross-entropy loss between student model prediction and ground truth, while the distillation loss is the Kullback-Leibler divergence loss between student model and teacher model prediction. The training process optimizes the final loss which is a combination of weighted distillation loss and student loss. Applying the knowledge distillation method, the team has reduced the model parameter of the ensemble models 9 times with the prediction performance declining only 0.4% (experiment results are shown in Table 5).

In addition, to improve the performance of the classification models, some other deep learning methods are applied, including data augmentation, up-sampling positive labels, pseudo labels, and class-weight technique. These methods help solving the imbalanced data problem that is critical in X-ray diagnosis. The imbalance data problem happens when a dataset consists of many "normal" samples with a small percentage of "abnormal" ones. It can lead to the machine learning models' bias toward the majority class. Besides, the auto noise label correction method is utilized for data cleaning, whereby suboptimal images can be re-corrected by a trained model. Importantly, some deep learning-based techniques such as using attention networks[23] and test time argumentation TTA[24] are also efficiently applied to improve model accuracy.

The TB X-ray image segmentation is a critical task in labeling each pixel of an abnormal TB finding in CXR images for detecting and segmenting lesions. The lesion segmentation helps doctors identify a particular area of the disease and extract detailed information for a more accurate diagnosis. For TB area segmentation task, the team has developed a model based on the Unet. After experiments on many different encoder networks, the Efficient-net B5 model is the most appropriate network for the backbone of the encoder site. The training process is done in five iterations. For each iteration, training time is set to 30 epochs with 2 NVDIA Tesla V100 GPUs. The TB segmentation process is described in Figure 13.

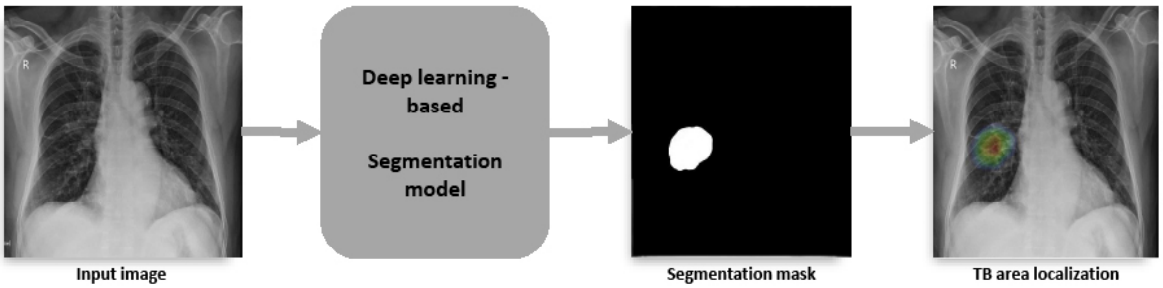

Figure 13: Tuberculosis area segmentation approach

5.4 Model evaluation

The trained model is evaluated on the test set to measure the model quality. The output of this step is a set of metrics to assess the quality of the model. The team has conducted experiments to test the models on the pre-defined test set for both classification and segmentation tasks.

The binary classification model is evaluated using different metrics such as F1 Score, AUC, Sensitivity, and Specificity. Table 6 presents the efficiency of using the Knowledge Distillation (KD) method to optimize the model. The KD method has dropped the prediction performance from 0.838 to 0.834 F1-Score, but the model parameter has been reduced nine times from 278M to 30M when replacing the ensemble models with a single model.

Table 6. The efficient result of Knowledge Distillation method

Model	AUC	F1 Score	Model Parameter
Ensemble model	0.970	0.838	278M
Single model Efficient-net B5 without KD	0.945	0.817	30M
Efficient-net B5 with KD	0.968	0.834	30M

The quality* of trained classification model on the test set of 9,437 images is AUC=0.968, F1 score=0.834, Sensitivity=0.86 and Specificity=0.961 with selected cut-off value=0.35. The performance of the model is also shown via the receiver operating characteristic (ROC) curve in Figure 14.

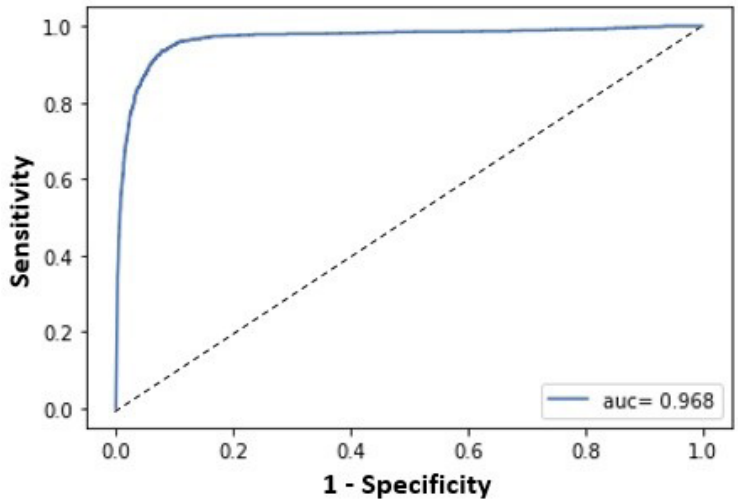

Figure 14: The ROC of TB classification model

The test result proves the reliability of the trained model on the TB classification task.

For the TB area segmentation task, the model quality is evaluated by comparing segmentation masks of model prediction with doctors' annotation masks on pixel level. A Dice score metric is selected to evaluate segmentation models. Our model achieves 0.753 Dice score* on the test set. Even when there is no existing approach for TB segmentation on X-ray images, when comparing with another method for TB segmentation on CT image[25] with 0.74 dice score, the result shows that the segmentation model has performed well.

** This is the performance of AI model v1 released on January 5, 2021*

5.5 Model deployment and monitoring

To deploy a TB model, in reality, there needs to be a transfer gate to connect through the software platform to support doctors. The research team decided to integrate such platform into DrAid™; by doing this, the AI model can diagnose TB disease simultaneously with other lung diseases. Initially, the X-ray images will be directly transferred from X-ray machines through an intermediary system for pre-processed images. The primary purpose of this system is to delete patient identifying information, examine the quality, and choose the qualified images for the TB model. After verification, all qualified images will be sent through VinBrain Azure Cloud. The results will be presented including the probability of TB disease, and highlight any areas affected by the disease. Using DrAid™ platform, doctors can interact with the TB model by sending images from X-ray machine and receiving diagnosis results from WebApp platform.

To maintain the speed and sustainability between the Cloud system environment and low power devices from end-users, the PyTorch model has been leveraged to maximize the efficiency of CPUs. The process to deploy the model also requires supervision from different versions of the model and sources code that are implemented. Therefore, GIT[26] and DVC tools[27] will assist this task most efficiently. Besides, Jenkins[28] tool will assist the automatization of the construction, examination, and execution of the model based on pre-build protocols with the lowest rate and risks possible.

In reality, the TB model will be developed, improved and deployed on a regular basis. To check and detect all problems and errors that affect the results creation process, the Azure Cloud tool help store all issues and errors encountered. As a result, the engineer can analyze and provide the best solutions to improve and prevent any delays in the future.

5.6 About VinBrain: the first AI for healthcare company in Vietnam

VinBrain is the leading AI for Healthcare products company in Vietnam, funded by Vingroup - the largest conglomerate by market capitalization in Vietnam. Our mission is to infuse AI and IoT into healthcare to improve human lives and productivity. VinBrain aims to provide equitable access for everyone to the best healthcare solutions, knowledge, and services. The company has attracted hundreds of talented applied scientists, software engineers, and product managers with world-class experience and expertise in machine learning, computer vision, NLP, recommender system and large-scale products and services development.

DrAid™ for Radiology (Figure 14), developed by VinBrain, greatly contributes to helping doctors in the diagnosis of 21 abnormalities and diseases with an average F1-score of 89.3% (no finding, COVID-19, pulmonary tuberculosis, pneumonia, pleural effusion, pleural other, neumothorax, lung opacity, atelectasis, consolidation, edema, pulmonary scar, lung lesion, nodule, mass, cavitation, enlarged cardiomediastinum, cardiomegaly, widening mediastinum, fracture, medical device).

In addition, DrAid™ automatically generates medical reports, including borders, heat maps, and accurate measurements of the abnormal areas. Besides, to optimize the process of making report, DrAid™ has a Speech-to-Text feature, allowing doctors to edit the report using their voice.

Figure 15: Doctor uses DrAid™, a product of VinBrain

Currently DrAid™ has been deployed in 84 hospitals and being actively used by 493 doctors in Vietnam. Until now, more than 233,000 medical images have been diagnosed with DrAid™'s help.

6. Next steps

Following the rapid progress of DrAid™'s development, the product's performance will be independently assessed in the next steps. This validation process will include a series of benchmarking exercises, followed by a prospective verification study.

The benchmarking exercise will employ well-characterized test libraries constituted using CXR images from Friends for International TB Relief's (FIT) community-based, mobile CXR screening campaigns and with generous support from the Stop TB Partnership. The libraries will contain a broad set of demographic and clinical parameters, which will enable the performance assessment of DrAid™ across a broad spectrum of factors against both human CXR readers and other CAD software solutions. Furthermore, it will inform VinBrain of a possible set of thresholds for field implementation. One of FIT's test libraries was recently used in a comparative landscape evaluation of 12 other CAD software solutions for TB screening.

The subsequent verification study will consist of deploying the tool in the field alongside a blinded field radiologist for community-based active TB case finding activities as well as facility-based surveillance of CXR for suspected TB. In these two settings, DrAid™ will read all CXRs in parallel with human doctors with a comparison of sensitivity and specificity performed based on patients with active TB detected. Active TB cases will include both patients who were solely confirmed through a positive sputum test as well as patients, whose diagnosis were based on either bacteriologic confirmation or clinical suspicion.

To ensure scientific excellence in the implementation of the performance evaluation of DrAid™, VinBrain and FIT are receiving additional technical support from experts of the US CDC's Vietnam country office. This support includes review and guidance to promote adherence to standard operating procedures and ensure high analytical rigor and ethical standards in the implementation of the evaluation.

This convergence of industry, non-profit and bilateral technical agency represents a critical innovation in public health in general and specifically in the field of TB care and prevention. If carried out successfully, this project has the potential to yield both a high-quality product for improving TB case finding in Vietnam and worldwide as well as furnishing a blueprint for pathfinding, trilateral collaboration model in health.

References

1. World Health Organization Global Tuberculosis Report 2019; Geneva, Switzerland, 2019;

2. Office of the Prime Minister Approval of the National Strategy for TB prevention and control until 2020 with vision to 2030 [vietnamese]; Viet Nam, 2014;

3. The second national tuberculosis prevalence survey in Vietnam; 2020;

4. Program, V.N.N.T. Measuring the level of under-reporting and estimating incidence for tuberculosis in Viet Nam: Preliminary Results TB Inventory Study 2016-2017. In Proceedings of the Workshop Report January; Ha Noi, Vietnam, 2018

5. Nhung, N. V.; Hoa, N.B.; Anh, N.T.; Anh, L.T.N.; Siroka, A.; Lönnroth, K.; Garcia Baena, I. Measuring catastrophic costs due to tuberculosis in Viet Nam. Int. J. Tuberc. Lung Dis. 2018, 22, 983–990, doi:10.5588/ijtld.17.0859.;

6. Hoa, N.B.; Cobelens, F.G.J.; Sy, D.N.; Nhung, N. V.; Borgdorff, M.W.; Tiemersma, E.W. First national tuberculin survey in Viet Nam: Characteristics and association with tuberculosis prevalence. Int. J. Tuberc. Lung Dis. 2013, 17, 738–744, doi:10.5588/ijtld.12.0200;

7. Dye, C.; Glaziou, P.; Floyd, K.; Raviglione, M. Prospects for Tuberculosis Elimination. Annu. Rev. Public Health 2013, 34, 271–286, doi:10.1146/annurev-publhealth-031912-114431;

8. Hoa, N.B.; Sy, D.N.; Nhung, N.V.; Tiemersma, E.W.; Borgdorff, M.W.; Cobelens, F.G. National survey of tuberculosis prevalence in Viet Nam. Bull. World Health Organ. 2010, 88, 273–280, doi:10.2471/BLT.09.067801.;

9. Boehme, C.C.; Nabeta, P.; Hillemann, D.; Nicol, M.P.; Shenai, S.; Krapp, F.; Allen, J.; Tahirli, R.; Blakemore, R.; Rustomjee, R.; et al. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. N. Engl. J. Med. 2010, 363, 1006–15, doi:10.1056/NEJMoa0907847;

10. World Health Organization; Automated real-time Nucleic Acid Amplification Technology for rapid and simultaneous detection of Tuberculosis and Rifamcin resistance: Xpert MTB / Rif system policy statement; 2011;

11. World Health Organization Automated Real-Time Nucleic Acid Amplification Technology for Rapid and Simultaneous Detection of Tuberculosis and Rifampicin Resistance: Xpert MTB/RIF Assay for the Diagnosis of Pulmonary and Extrapulmonary TB in Adults and Children: Policy update; Geneva, 2013;

12. Theron, G.; Peter, J.; Dowdy, D.; Langley, I.; Squire, S.B.; Dheda, K. Do high rates of empirical treatment undermine the potential effect of new diagnostic tests for tuberculosis in high-burden settings? Lancet Infect. Dis. 2014, 14, 527–532, doi:10.1016/S1473-3099(13)70360-8.;

13. WHO consolidated guidelines on tuberculosis Module 2: Screening – Systematic screening for tuberculosis disease, 2021;

14. Pyimagesearch (https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-objectdetection)

15. Guidelines on TB Diagnosis, Treatment and Prevention (1314/QĐ-BYT); Viet Nam Ministry of Health: Ha Noi, Viet Nam, 2020.

16. Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation" in MICCAI 2015;

17. Stone M. Cross-validatory choice and assessment of statistical predictions. J. Royal Stat. Soc., 36(2), 111–147, 1974.;

18. Cawley, Gavin C.; Talbot, Nicola L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. 11. Journal of Machine Learning Research: 2079– 2107, 2010.;

19. Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, "Densely Connected Convolutional Networks" in CVPR 2017;

20. François Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions" in CVPR 2017.;

21. Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, "Aggregated Residual Transformations for Deep Neural Networks" in CVPR 2017.;

22. Mingxing Tan, Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" in ICML 2019.

23. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio, "Show, Attend and Tell: Neu-

ral Image Caption Generation with Visual Attention" Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:2048-2057, 2015;

24. Divya Shanmugam, Davis Blalock, Guha Balakrishnan, John Guttag, "When and Why Test-Time Augmentation Works" in arXiv:2011.11156;

25. Pedro M. Gordaliza; Arrate Muñoz-Barrutia; Mónica Abella; Manuel Desco; Sally Sharpe & Juan José Vaquero "Unsupervised CT Lung Image Segmentation of a Mycobacterium Tuberculosis Infection Model" in Scientific Reports volume 8, Article number: 9802 (2018)

26. Github (https://github.com/microsoft/onnxruntime)

27. DVC (https://dvc.org/)

28. Jenkins (https://www.jenkins.io/)

For further information, please contact:
VINBRAIN LLC (VINGROUP JSC)
No. 7, Bang Lang 1 Street, Vinhomes Riverside,
Viet Hung Ward, Long Bien District, Ha Noi, Viet Nam
Tel: (+84) 24 7106 8680
Email: info@vinbrain.net
Website: www.vinbrain.net

For further information, please contact:
Freundeskreis für Internationale Tuberkulosehilfe e.V (FIT)
No. 1, Alley 21, Le Van Luong Street,
Nhan Chinh Ward, Thanh Xuan District, Ha Noi, Viet Nam
Tel: (+84) 24 7300 0084
Email: info@tbhelp.org
Website: www.tbhelp.org