


## RESEARCH ARTICLE

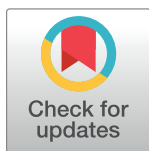
# Learning to diagnose common thorax diseases on chest radiographs from radiology reports in Vietnamese

Thao Nguyen<sup>1</sup> , Tam M. Vo<sup>1</sup> , Thang V. Nguyen<sup>1</sup>, Hieu H. Pham<sup>1,2,3\*</sup> , Ha Q. Nguyen<sup>1,2</sup>

**1** Smart Health Center, VinBigData JSC, Hanoi, Vietnam, **2** College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam, **3** VinUni-Illinois Smart Health Center, Hanoi, Vietnam

 These authors contributed equally to this work.

\* [hieu.ph@vinuni.edu.vn](mailto:hieu.ph@vinuni.edu.vn)



## Abstract

Deep learning, in recent times, has made remarkable strides when it comes to impressive performance for many tasks, including medical image processing. One of the contributing factors to these advancements is the emergence of large medical image datasets. However, it is exceedingly expensive and time-consuming to construct a large and trustworthy medical dataset; hence, there has been multiple research leveraging medical reports to automatically extract labels for data. The majority of this labor, however, is performed in English. In this work, we propose a data collecting and annotation pipeline that extracts information from Vietnamese radiology reports to provide accurate labels for chest X-ray (CXR) images. This can benefit Vietnamese radiologists and clinicians by annotating data that closely match their endemic diagnosis categories which may vary from country to country. To assess the efficacy of the proposed labeling technique, we built a CXR dataset containing 9,752 studies and evaluated our pipeline using a subset of this dataset. With an F1-score of at least 0.9923, the evaluation demonstrates that our labeling tool performs precisely and consistently across all classes. After building the dataset, we train deep learning models that leverage knowledge transferred from large public CXR datasets. We employ a variety of loss functions to overcome the curse of imbalanced multi-label datasets and conduct experiments with various model architectures to select the one that delivers the best performance. Our best model (CheXpert-pretrained EfficientNet-B2) yields an F1-score of 0.6989 (95% CI 0.6740, 0.7240), AUC of 0.7912, sensitivity of 0.7064 and specificity of 0.8760 for the abnormal diagnosis in general. Finally, we demonstrate that our coarse classification (based on five specific locations of abnormalities) yields comparable results to fine classification (twelve pathologies) on the benchmark CheXpert dataset for general anomaly detection while delivering better performance in terms of the average performance of all classes.

## OPEN ACCESS

**Citation:** Nguyen T, Vo TM, Nguyen TV, Pham HH, Nguyen HQ (2022) Learning to diagnose common thorax diseases on chest radiographs from radiology reports in Vietnamese. PLoS ONE 17(10): e0276545. <https://doi.org/10.1371/journal.pone.0276545>

**Editor:** Tarik A. Rashid, University of Kurdistan Hewler, IRAQ

**Received:** May 12, 2022

**Accepted:** October 7, 2022

**Published:** October 31, 2022

**Copyright:** © 2022 Nguyen et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Data are available from the Institutional Review Board (IRB) of the Phu Tho General Hospital. Data access may be requested from Dr. Luc Quang Nguyen, Head of Radiology Department, Phu Tho General Hospital, at "[drtranquangluc@gmail.com](mailto:drtranquangluc@gmail.com)," for researchers who meet the criteria for access to confidential data.

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

Radiography has always been one of the most ubiquitous diagnostic imaging modalities so far, while chest X-ray (CXR) is the most commonly performed diagnostic X-ray examination [1]. CXRs has an important role in clinical practice, effectively assisting radiologists to detect pathologies related to the airways, pulmonary parenchyma, vessels, mediastinum, heart, pleura and chest wall [2]. In recent years, great advances in GPU computing and research in the fields of machine learning have led to the trend of automating CXR image diagnostics [3–9] and many other X-ray modalities [10–13]. In addition, the availability of large-scale public datasets [14–19] has sparked interest in study and application, with some of them already being used and integrated into the Computer-Aided Diagnosis (CAD) system to reduce the rate of CXR misdiagnosis.

Several datasets, including CheXpert [14], MIMIC-CXR [15], PadChest [16], Chest-xray8, Chest-xray14 [17] and VinDr-CXR [19, 20], VinDr-PCXR [21, 22], had a significant impact on increasing labeling methods and model quality. Building a reliable CXR dataset for a specific project, on the other hand, remains a difficult and challenging task because medical data is difficult to obtain due to numerous restrictions on patient information confidentiality, and label quality is heavily influenced by the doctors' experience and subjective opinion [1]. This is costly and time-consuming but essential, especially for a task that tackles specific challenges, such as focusing on a certain set of patients or illnesses. In such a way that adopting the afore-said large-scale datasets is sometimes ineffective, possibly because the image quality, labeling, or data characteristics are no longer appropriate. Additionally, CXR images and medical reports corresponding to each examination are also stored in hospital storage systems such as Picture Archiving and Communication System (PACS) and Hospital Information System (HIS) during the radiology process. This is a tremendous available resource to build large-scale CXR datasets in which the annotation can be automatically interpolated from the free text report without any involvement of radiologists. Therefore, pipelines or methods to create datasets from available resources are always valuable.

Some previous works also developed methods to relabel public large datasets or constructed a new one. Wang et al. [17] proposed a method for extracting a hospital-scale CXR dataset from the PACS via an unified weakly-supervised multi-label image classification and disease localization formulation by applying natural language processing (NLP) techniques. NegBio [23], a rule-based algorithm that utilizes universal dependencies and subgraph matching, known as providing regular expression infrastructure for negation and uncertain detection in radiology reports. Filice et al. [24] investigated the benefit of utilizing AI models to create annotations for review before adjudication in order to speed up the annotation process while sacrificing specificity. Johnson et al. [15] extracted and classified mentions from the associated reports using two NLP tools, CheXpert and NegBio, before aggregating them to arrive at the final label. To construct structured labels for the images, Irvin et al. [14] created an automated rule-based labeler to extract observations and capture uncertainties contained in free-text radiology reports. Padchest [16] labeled the majority of the dataset using a recurrent neural network with an attention mechanism. This dataset contains excerpts from Spanish radiology reports, however the labels have been mapped to biological vocabulary unique identifier codes, making the resource useful regardless of the language. RadGraph [25] introduced a new dataset of clinical entities and relations annotated in full-text radiology reports taken from CheXpert and MIMIC. This research made use of a novel information extraction schema that extracts clinically relevant information associated with a radiologist's interpretation of a medical image.

More advanced NLP approaches, such as Bidirectional Encoder Representations from Transformers (BERT) [26], are used in some studies. Chexpert++ [27], a BERT-based, high fidelity approximation labeler applied to CheXpert, is significantly faster, fully differentiable, and probabilistic in outputs. VisualCheXbert [28] utilized a biomedically-pretrained BERT model to map directly from a radiology report to the image labels, with a supervisory signal determined by a computer vision model trained to detect medical conditions from chest X-ray images. CheXbert [29] is a BERT-based approach to medical image report labeling that exploits both the scale of available rule-based systems and the quality of expert annotations.

Dictionary-based heuristics are another popular way for creating structured labels from free-text data. For instance, MedLEE [30] utilizes a pre-defined lexicon to convert radiology reports into a structured format. Mayo Clinic's Text Analysis and Knowledge Extraction System (cTAKES) [31] tool combines dictionary and machine learning methods, and uses the Unified Medical Language System <https://www.nlm.nih.gov/research/umls/index.html> (UMLS) for dictionary inquiries. Dictionary-based NLP systems have a key flaw is that they do not always establish high performance when handling in-house raw clinical texts, especially those with misspellings, abbreviations, and non-standard terminology. On top of that, the mentioned systems only cover English language and cannot handle non-English clinical texts. Languages other than English, including Vietnamese, do not have sufficient clinical materials to build a medical lexicon. In nations where English is not the official language, this has been a huge obstacle in building clinical NLP systems. In current work, our data pipeline can be applied for the available data in PACS and HIS, which can assist minimize data labeling costs, time, and effort while reducing radiologists' involvement in the workflow. We propose a set of matching rules to convert a typical radiology report to the normal/abnormal status of classes.

Other than the above-mentioned differences in labeling methods, our label selection is also different from previous studies. So far, most of the studies were developed for classifying common thoracic pathologies or localizing multiple classes of lesions. For instance, most deep learning models were developed on the MIMIC-CXR [32] and CheXpert [33–35] datasets for classifying 14 common thoracic pathologies on CXRs in recent years. The earlier dataset ChestX-ray14 [17], an expansion of ChestX-ray8 [17], including the same set of 14 findings has been used to develop deep learning models [36, 37]. Nevertheless, these approaches are far different from how Vietnamese radiologists work. In clinical practice, a CXR radiology report always includes four descriptions that correlate to four fixed anatomical regions of the thorax: chest wall, pleura, pulmonary parenchyma and cardiac. Therefore, it is not practical for Vietnamese radiologists to utilize a CAD system that provides suggestions for the presence of 14 diseases. Typically, when examining a CXR image, radiologists analyze that image by region; consequently, it is more convenient for the system to indicate the abnormality of each area, eliminating the need to match the lesion type with the region being viewed. To address the realistic demand of Vietnamese radiologists, we developed a system to classify CXRs into 5 classes depending on the position of pathologies: chest wall, pleura, parenchyma, cardiac abnormality and the existence of abnormalities in the CXRs, if any. When tested on the benchmark CheXpert dataset, we found that this coarse classification produces results comparable to the detailed classifier of 14 findings in terms of abnormal class and gives better results in terms of macro average F1 score of all classes.

Our work was developed on the dataset collected at Phu Tho General Hospital—a Vietnamese provincial hospital. To develop trainable images with corresponding labels, DICOM files in PACS are matched with radiology reports retrieved from HIS. By extracting data from radiology reports, generating normal/abnormal status of 5 classes and treating it as the ground-truth reference, we can conclude that there were positive results when classifying CXRs according to 5 groups of pathologies, which are modeled after the radiologist's description in

their medical report. Unlike the automatic data labeling methods mentioned above, our proposed method is simple yet accurate by filtering the descriptions alluding to no findings first, then searching for phrases implying abnormalities in each position. Therefore, the labeling process is strictly controlled through stages, making it easy to detect errors and correct them. In addition, adding a manual step to the labeling process helps us deal with misspellings, which was neglected by the previous method. In this step, we also find infrequent phrases, adding them to our list of phrases indicating abnormality to make it more complete. Furthermore, a report always includes descriptions corresponding to four fixed anatomical regions of the thorax, thus by generating set of labels matching these regions, we can minimize the chance that a label is uncertain.

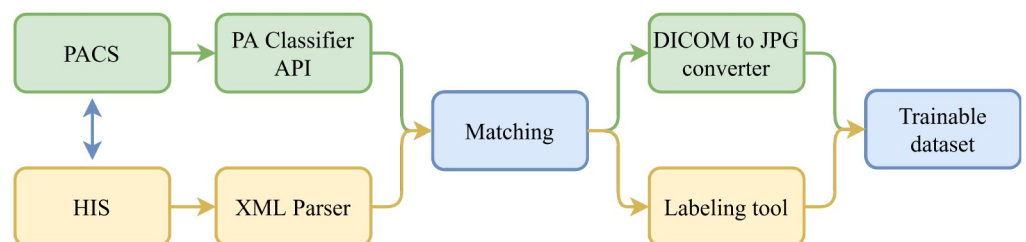
## Material and method

### Dataset building pipeline

Our proposed pipeline consists of five steps: (1) data collection, (2) PA-view filtering, (3) XML parser, (4) data matching and (5) data annotation. Fig 1 illustrates the above five steps in detail. Firstly, DICOM files stored in PACS will be acquired and filtered to retain only posterior-anterior (PA) view CXRs by the PA classifier application programming interface (API). Meanwhile, radiology reports stored in HIS as XML files will be parsed to attain some specific information. Afterward, DICOM files and radiology reports belonging to the same patient will be matched to generate pairs of DICOM-XML files of the same examination. Once a DICOM file has been determined to match with an XML file, that DICOM file will be converted to JPG format and the XML file will be the subject of a labeling tool to generate a set of corresponding labels. At the end of the procedure, we can obtain a trainable dataset which includes JPG images and their corresponding labels.

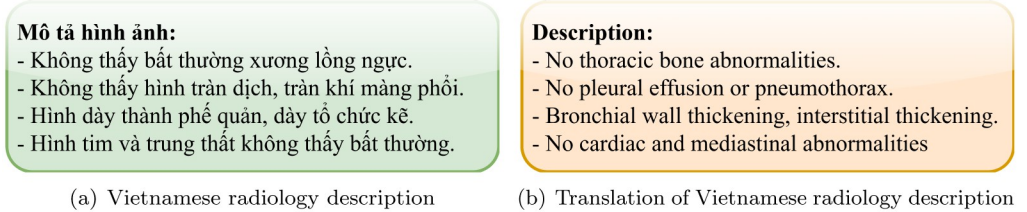
**Data collection.** We retrospectively collected chest radiography studies from Phu Tho General Hospital, which were performed within five months from November 2020 to March 2021, along with their associated radiology reports. The ethical clearance of these studies was approved by the Institutional Review Board (IRB) of Phu Tho General Hospital. With this approval, the IRB allows us to access their data and analyze raw chest X-ray scans using our VinDr's platform, which will be used for data filtering. The need for obtaining informed patient consent was waived because this retrospective study did not impact clinical care or workflow at the hospitals, and all patient-identifiable information in the data has been removed.

We decided to select four types of pathologies because of their prevalence in the medical reports and clinical practice. An example of a typical description extracted from a radiology report is shown in Fig 2. The description is divided into four main categories: lungs, cardiac, pleura and chest wall by most Vietnamese radiologists. From the four groups of pathology, we



**Fig 1. Overview diagram of the process of collecting and building medical image dataset.** The process consists of five steps: data collection from PACS and HIS, PA-view filtering, XML parser, data matching and data annotation.

<https://doi.org/10.1371/journal.pone.0276545.g001>



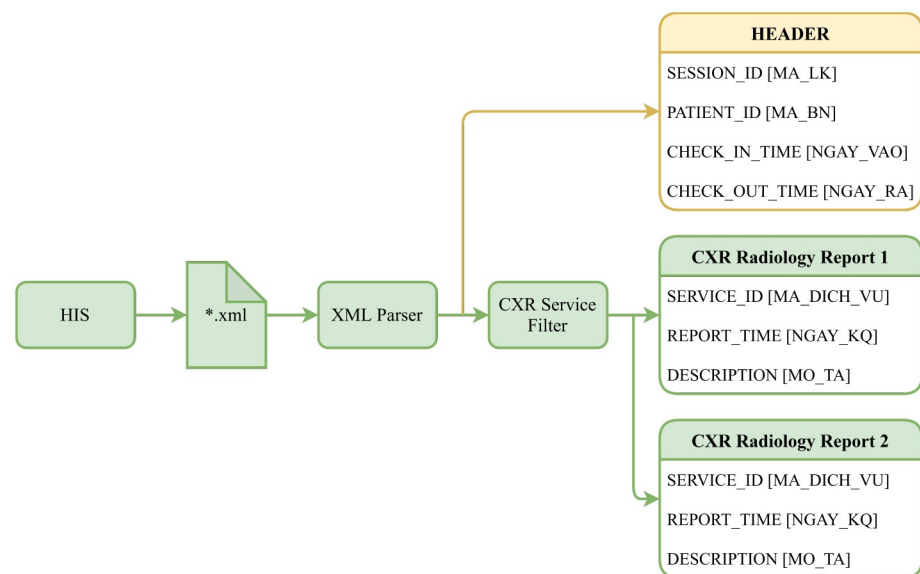
**Fig 2. The description in a typical radiology report in Vietnam.** The description is divided into four main categories: chest wall, pleura, lungs (parenchyma) and cardiac.

<https://doi.org/10.1371/journal.pone.0276545.g002>

create an annotation set consisting of five classes, with the first four classes corresponding to these four groups and the other indicating the presence of abnormalities on CXRs, if any.

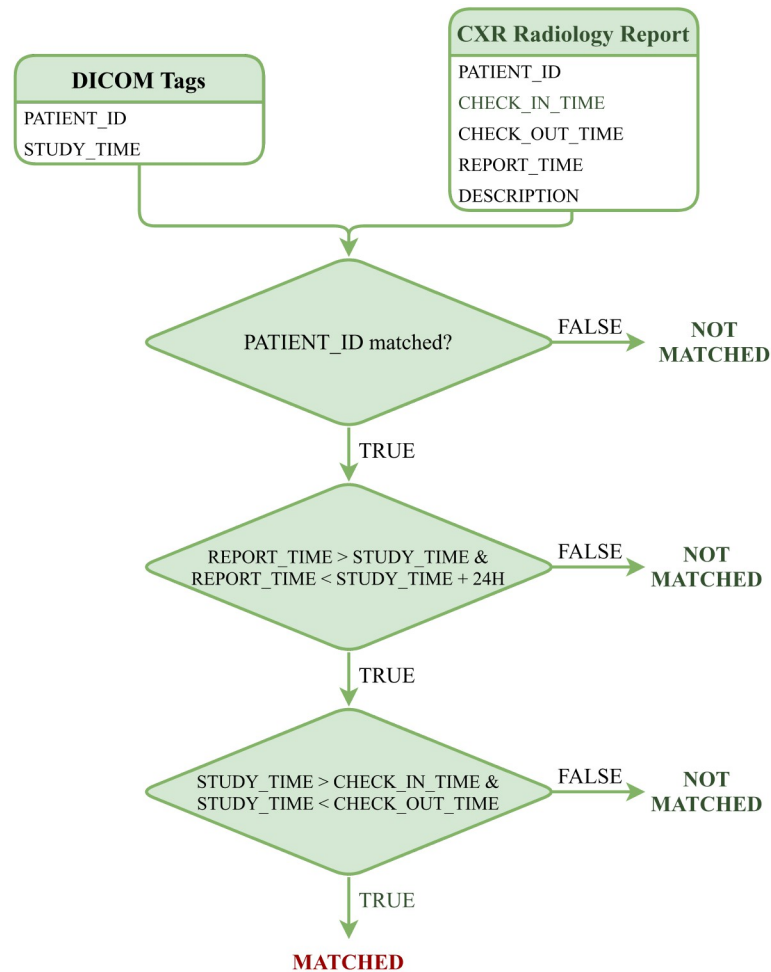
**PA-view filtering.** The collected data was mostly of Posterior-Anterior (PA)-view CXR, but also included a large number of outliers such as images of body parts other than chest, low-quality images or images with different views than PA-view. To guarantee that only CXRs of PA-view will be retained, we ran an API that is powered by VinDr’s platform <https://vindr.ai/vindr-lab>. The API takes a DICOM file as an input and returns the probability that the image saved in that file is a PA-view CXR. The DICOM file will proceed to the next stage of data pre-processing if this probability exceeds 0.5—a normalized threshold; else, the file will be marked as ignored.

**XML parser.** We use the same procedure for the XML parsing and data matching process as in our previous study [38], shown in Fig 3. The figure illustrates the procedure of extracting radiology reports from HIS. Each assessment and treatment session was saved in the Extensible Markup Language (XML) file format by HIS. A session includes all information of the patient between check-in and check-out time. The XML parser can read the header of a session that includes SESSION\_ID, PATIENT\_ID, CHECK\_IN\_TIME, and CHECK\_OUT\_TIME. These attributes are shared among all radiology reports belonging to the same session and will



**Fig 3. Radiology reports extraction process for CXR examinations collected from HIS [38].** The original Vietnamese counterparts are put inside square brackets.

<https://doi.org/10.1371/journal.pone.0276545.g003>



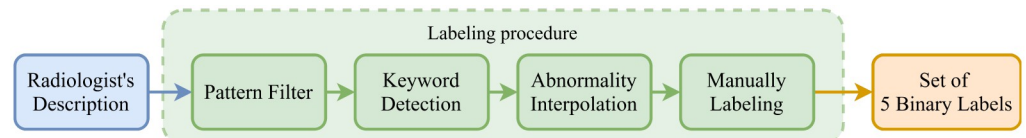
**Fig 4. Algorithm for matching a DICOM file obtained from PACS with a radiology report collected from HIS.**

<https://doi.org/10.1371/journal.pone.0276545.g004>

be used to link to the corresponding DICOM file. All reports are also interpreted using the XML parser to obtain the SERVICE\_ID, REPORT\_TIME, and DESCRIPTION properties. Only reports with a SERVICE ID matching the values expressly assigned by the Vietnamese Ministry of Health for chest radiography were preserved to exclude extraneous reports.

**Data matching.** To match the DICOM file with the corresponding XML file, we have simulated the algorithm in [38], which is depicted in Fig 4. Since the HIS and PACS are linked by PATIENT\_ID, this key is used by the matching algorithm to determine whether the DICOM file and radiography report belong to the same patient. Moreover, REPORT\_TIME must be within 24 hours of STUDY\_TIME, which is a regulated protocol of the hospital. Finally, STUDY\_TIME has to be between CHECK\_IN\_TIME and CHECK\_OUT\_TIME. If all of the conditions are fulfilled, the DICOM file and the radiology report are matched.

One problem we encountered here is that one DICOM file matched multiple reports and vice versa, because their STUDY\_TIME attributes were separated by a period of less than 24 hours. In such a short period of time, the examination results are often the same, the reason for taking additional radiographs may be due to the poor quality of the first image. Therefore, the description from the reports is usually the same, and this DICOM file is assigned to one of



**Fig 5. Semi-automated data annotation pipeline.** The system consists of 4 steps, the first 3 steps are automatic and the last one is carried out manually.

<https://doi.org/10.1371/journal.pone.0276545.g005>

the matched reports. In several cases where the descriptions in the reports are different, the DICOM file will be given to a radiologist to review and match the correct report.

**Data annotation.** After extracting descriptions that match the DICOM files, we developed a simple labeling algorithm that takes the radiologists' description as input and returns a list of five binary elements, corresponding to the presence or absence of abnormalities belonging to 5 classes. Fig 5 illustrates the major steps of data annotation, which is implemented in semi-automated manner, including (1) pattern filtering, (2) keyword detection (3) abnormality interpolating and (4) manually labeling.

**Pattern Filtering** The dataset we obtained from Phu Tho General Hospital is unbalanced, with the majority of the images exhibiting no pathology. We have obtained 1,568 different templates from all the descriptions. Filtering descriptions that are elements of the predetermined set of templates (specifically 11 templates imply no findings) would help us save a significant amount of time when it comes to data labeling. A CXR is considered normal if one of 11 templates exactly appears in the DESCRIPTION of the corresponding radiology report.

**Keyword detection** After pattern filtering, most of the instances without pathologies are retained. In this step, we have to handle most of the abnormality descriptions and some remaining normality ones. Keyword detection is divided into four sub-stages, which could be performed simultaneously, to detect keywords indicating abnormalities in the chest wall, pleura, parenchyma, and mediastinum. To find keywords for each class, e.g. chest wall, we break down the radiologist's description into 4 categories (categories are separated by "-" (dash) in the radiology descriptions). From the sentences in the chest wall category, we gather keywords indicating abnormalities, such as "fracture", "osteoporosis", "bone fusion surgery" to create the fixed set of keywords. Descriptions containing keywords in the chest wall set will be annotated as 1 for the corresponding class, similarly for the pleura, parenchyma, and cardio classes. Some common keywords setting for the four classes are listed in Table 1.

**Table 1. Examples of Vietnamese keywords indicate abnormalities in chest wall, pleura, parenchyma, cardiac classes and abnormality out of these four group.** English translations are enclosed in square brackets.

Class name	Keywords
Chest wall (bone)	Gãy xương [Bone fracture] Thưa xương [Osteoporosis] Tiêu xương [Bone resorption]
Pleura	Dày màng phổi trái/phải [Left/right pleural thickening] Mờ góc sườn hoành màng phổi trái/phải [Left/right costophrenic angle blunting] Tù góc sườn hoành trái/phải [Loss of the left/right costophrenic angle]
Parenchyma	Dày thành phế quản [Bronchial wall thickening] Dày tổ chức kẽ [Interstitial pulmonary thickening] Đài mờ giữa phổi trái/phải [Opacity between left/right lung]
Cardio	Quai động mạch chủ (đmc) vồng [Ascending aortic arch] Hình tim trái/phải to [Enlarged /right cardiomegaly] Giãn cung thất trái/ phải [Left/right ventricular arch dilatation]
Other abnormality	Liềm hơi dưới vòm hoành trái/phải [Sickle of air below the left/right diaphragm]

<https://doi.org/10.1371/journal.pone.0276545.t001>

Table 2. Number of instances which contain five labeled observations in training, validation and the whole dataset.

Position of pathology	Positive		Negative	
Chest wall	Training	166	Training	6835
	Validation	71	Validation	2930
	Total	237 (2.37%)	Total	9765 (97.63%)
Pleura	Training	155	Training	166
	Validation	67	Validation	71
	Total	222 (2.22%)	Total	9780 (97.78%)
Parenchyma	Training	1520	Training	6846
	Validation	652	Validation	2934
	Total	2172 (21.72%)	Total	7830 (78.28%)
Cardio	Training	548	Training	6453
	Validation	235	Validation	2766
	Total	783 (7.83%)	Total	9219 (92.17%)
Abnormal	Training	1976	Training	5025
	Validation	848	Validation	2153
	Total	2824 (28.23%)	Total	7178 (71.77%)

<https://doi.org/10.1371/journal.pone.0276545.t002>

**Abnormality interpolating** The first four classes have been annotated at the keyword detection stage, here the abnormality class labeling is implemented by inferring from those others. Abnormality value will be set to 1 (positive) if any of the other classes are noted as anomalies or has any other anomaly even though it does not belong to the four groups above.

**Manual Labeling** Descriptions that neither belong to the 11 normality templates nor contain any of the keywords in the four fixed sets have a high probability of being misspelled or describing rare pathologies or including pathologies that cannot be assigned to one of the four main regions. To handle such cases, we inspected them to correct spelling mistakes manually, then forwarded confusing descriptions to a radiologist of Phu Tho General Hospital for annotating. These cases account for less than 0.5% of the total descriptions, thus labeling the remain is not a time-consuming task, that minimizes the doctor's involvement in data labeling.

Over five months, we obtained the total number of 12,367 XML files and 12,376 DICOM files corresponding to 11,088 studies. 10,847 DICOM files were PA chest radiographs, and 10,002 of them matched with information extracted from XML files. Table 2 details the number of positive and negative samples of the five classes in the collected dataset. For model development, we split the dataset into training and validation sets with the ratio of 7/3 and one constraint is that the distribution of each class in training and validation sets is approximated to the distribution of the original dataset.

### Quality control

To ensure the quality of the dataset is guaranteed, we randomly take 5% of the data to inspect if there are any inappropriate images or labels that do not match the corresponding report. If any incorrectness is found, we will find out and correct it, then the 5% selection process is repeated until no more errors are detected. The inspection was carried out by a medical student majoring in radiology and was double checked by a radiologist of Phu Tho General Hospital.

### Labeler results

We evaluate the effectiveness of the proposed labeling procedure by manually labeling the samples and considering the result as the ground truth. F1-score will be used as the main metric to evaluate the quality of our labeling tool.



**Table 3. Evaluation results of proposed labeling tool.** Evaluation was performed on 3001 samples of the validation set.

Class	TP	FP	TN	FN	Precision	Recall	F1 score
Chest wall	71	0	2930	0	1	1	1
Pleura	67	1	2933	0	0.9853	1	0.9926
Parenchyma	652	1	2347	1	0.9985	0.9985	0.9985
Cardio	235	0	2766	0	1	1	1
Abnormal	848	0	2153	0	1	1	1

<https://doi.org/10.1371/journal.pone.0276545.t003>

**Evaluation set.** The reported evaluation set consists of 3001 radiology reports from 3001 instances—that totally overlap with the reports in the validation set. We manually annotated these radiology reports without access to additional patient information. We labeled whether there is any abnormality in chest wall, pleura, pulmonary parenchyma and cardio following a list of labeling conventions that was agreed upon ourselves. After we independently labeled each of the 3001 reports, disagreements were resolved by consensus discussion or radiologist’s consultation. The resultant annotation serves as ground truth on the report in evaluation set.

**Evaluation results.** After having the results as the radiologists’ annotation, combined with the set of labels generated by our method, the evaluation results of each class are listed in Table 3, with the metrics of precision, recall and F1 score. Overall, our labeling pipeline delivers the high values of F1 score in all classes, with the lowest figures of 0.9926 and 0.9985—being recorded in pleura and parenchyma classes, respectively. In chest wall, cardio and abnormal classes, our tool delivers the highest performance, without any mislabeled samples.

## Experiment and results

### Model development

Chest X-ray interpretation with deep learning methods usually relies on pre-trained models developed for ImageNet. Nevertheless, it was proved that architectures achieving remarkable accuracy on ImageNet are unlikely to give the same performance when experienced on the CheXpert dataset and the choice of model family deliver better improvement than image resizing within a family for medical imaging tasks [39]. We decided to choose the model family that has been proved to be highly efficient for CXR interpretation—ResNet50 [40], DenseNet121 [41], Inception-V3 [42] and EfficientNet-B2 [43]. We also leverage large public CXR datasets such as CheXpert to develop pre-trained models and compare the use of some benchmark chest X-ray datasets for transfer learning to ImageNet pre-trained models. Furthermore, the unbalance between classes has a negative impact on our dataset; for example, the chest wall class has a positive/negative ratio of 0.003. To address this problem, along with the conventional Binary Cross Entropy Loss (BCE), we used and assessed other loss functions established for multi-label imbalanced datasets, such as Asymmetric Loss (ASL) [44] and Distribution-balanced Loss (DBL) [45].

For each model architecture, we use the Adam optimizer (beta1 = 0.9, beta2 = 0.999 and learning rate = 1e-3), cooperating with Cosine annealing learning rate with gradual warm-up scheduler, a batch size of 16, three different loss functions: cross-entropy, distribution-balanced and asymmetric loss, image sizes of 768 and 1024.

Training was conducted on a Nvidia GTX 1080 with CUDA 10.1 and Intel Xeon CPU ES-2609. For one run of a specific model, we train for 160 epochs and evaluate each model every 413 gradient steps. Finally, checkpoint with the highest F1-score will be considered the best model for each training procedure.

We also used the nonparametric bootstrap [46] to estimate 95% confidence intervals for each statistic. There are 3,000 replicates are drawn from the validation set, and the statistic is calculated for each replicate. This procedure generates a distribution for each statistic, by reporting the 2.5 and 97.5 percentiles, the confidence intervals are obtained and significance is assessed at the  $p = 0.05$  level.

## Experimental result

In this work, chest X-ray classification models were trained on the training set detailed in Table 2. The models are distinguished from each other based on four attributes: (1) model architecture, (2) pre-trained dataset, (3) loss function and (4) image size, while sharing the common training procedure. First, we compare the effect of using pre-trained datasets and the impact of some loss functions on the multi-label problem. We choose ImageNet and CheXpert to transfer their knowledge to our target data. BCE—a common loss function, ASL and DBL—the two loss functions for multi-label issue were used in our experiment. The reported metrics are macro average (Av.) F1-score, AUC, sensitivity and specificity of the five classes. We only use ResNet50 architecture to compare these aspects with the same setup hyper parameters.

As we can see in Table 4, model using ASL and CheXpert dataset as pre-trained-initial parameters give the best result. All the metrics are higher than that of the others, especially when using ASL. This loss function always gives big value but is very effective because it heavily “penalizes” misclassified positive samples and hardly penalizes easy negative one. CheXpert is also useful in spite of containing similar patterns to our target data. We decide to use pre-trained model by CheXpert and ASL for later experiments.

To discover which family of architectures really fits our dataset, we do more experiments with Inception-V3, DenseNet121 and EfficientNet-B2, which are reported to perform well with radiographic images; and two sizes of image 768 and 1024. The result is shown in Table 5, which indicates that bigger image sizes do not give rise to better results, but affect training time. In the matter of model architectures, EfficientNet-B2 outperforms the others. In conclusion, model with EfficientNet-B2 architecture and input size of 768 delivers the best performance.

Detailed result of our best model is also presented in Table 6. By using ASL, the chest wall class has improved significantly when increasing to nearly 32% compared to the model using BCE and not using CheXpert as pre-trained. The pleura class has less samples than the chest wall, but the results do not improve much after using ASL, possibly because the chest wall class has a more diverse number of abnormal manifestations in our data, so the model focused more on this class.

Fig 6 illustrates plots on all tasks. The model achieves the best AUC on pleura class (0.96), and the worst on chest wall class (0.81). The abnormal class recorded 0.87 AUC, the parenchyma and cardiac classes witness figures of 0.86 and 0.92, respectively.

The same procedure is also applied to build the two models of fine classification (detection of 14 pathologies) and coarse classification (detection of abnormalities in 4 locations in CXR images), in order to evaluate the effectiveness of the coarse classification compared to the fine classification. We use the CheXpert benchmark dataset to build and evaluate two models sharing the same configurations to retain the sense of objectivity. The data in the CheXpert dataset are labeled with 14 classes corresponding to 13 abnormalities in the chest radiograph and an implication of no findings. We infer where the lesion is in the 4 considered positions based on the type of lesion indicated in the CheXpert dataset. Table 7 shows the mappings between CheXpert data labels (14 classes) and the proposed set of labels (5 classes). Comparison of coarse and fine classification on Table 8. Based on the results shown in the Table 8, it can be

**Table 4. Experimental results with different pre-train datasets and loss functions.** Model pre-trained on CheXpert dataset and using Asymmetric loss function yields the best performance.

Pretrained dataset + Loss function	Class	F1 score	AUC	Sensitivity	Specificity
ImageNet + BCE	Bone	0.098	0.6622	0.3239	0.8713
	Pleura	0.4196	0.9348	0.4478	0.9843
	Parenchyma	0.5742	0.8351	0.6380	0.8378
	Cardio	0.4513	0.8605	0.5617	0.9212
	Abnormal	0.6366	0.8337	0.7323	0.7761
	<b>Average</b>	<b>0.4359</b>	<b>0.8253</b>	<b>0.5408</b>	<b>0.887</b>
ImageNet + ASL	Bone	0.3800	0.7123	0.2676	0.9966
	Pleura	0.4925	0.9239	0.4925	0.9884
	Parenchyma	0.5941	0.8389	0.5982	0.8846
	Cardio	0.5278	0.9115	0.6255	0.9367
	Abnormal	0.6674	0.8482	0.7123	0.8337
	<b>Average</b>	<b>0.5324</b>	<b>0.847</b>	<b>0.5392</b>	<b>0.928</b>
ImageNet + DBL	Bone	0.1882	0.6993	0.1010	0.9799
	Pleura	0.2647	0.8691	0.403	0.9625
	Parenchyma	0.5566	0.8195	0.6748	0.7918
	Cardio	0.3929	0.8289	0.4723	0.9208
	Abnormal	0.6123	0.8126	0.6993	0.7696
	<b>Average</b>	<b>0.4029</b>	<b>0.8059</b>	<b>0.4909</b>	<b>0.8894</b>
CheXpert + BCE	Bone	0.0706	0.5412	0.0423	0.9962
	Pleura	0.2623	0.8540	0.2388	0.9867
	Parenchyma	0.537	0.7921	0.6396	0.0794
	Cardio	0.3872	0.8205	0.4638	0.9208
	Abnormal	0.581	0.7789	0.6875	0.7325
	<b>Average</b>	<b>0.3676</b>	<b>0.7573</b>	<b>0.4144</b>	<b>0.886</b>
CheXpert + ASL	Bone	0.4348	0.7757	0.3521	0.9935
	Pleura	0.5323	0.9424	0.4925	0.9918
	Parenchyma	0.6274	0.8624	0.6702	0.8706
	Cardio	0.5536	0.9197	0.6043	0.9508
	Abnormal	0.6777	0.8658	0.7512	0.8165
	<b>Average</b>	<b>0.5651</b>	<b>0.8732</b>	<b>0.5741</b>	<b>0.9247</b>
CheXpert + DBL	Bone	0.1674	0.6912	0.2535	0.957
	Pleura	0.4698	0.9513	0.5224	0.984
	Parenchyma	0.5958	0.8450	0.6104	0.8782
	Cardio	0.5094	0.9009	0.5745	0.9422
	Abnormal	0.6498	0.8493	0.7134	0.8100
	<b>Average</b>	<b>0.4758</b>	<b>0.8475</b>	<b>0.5349</b>	<b>0.9143</b>

<https://doi.org/10.1371/journal.pone.0276545.t004>

seen that the coarse classification method gives a higher F1 score in both the abnormal class and the macro average F1 score.

We also plot Grad-CAMs [47] to give the visual explanations of how the model fulfil predictions. Fig 7 illustrates the original images and their respective Grad-CAMs. In both cases, the pathologies in the collarbone (nondisplaced fracture) and in the pleura (pleural effusion) were correctly highlighted. The results are attained when performing with the EfficientNet-B2 architecture, the input size is 768x768, using the CheXpert dataset to build the pretrained model and apply the asymmetric loss function.

**Table 5. Experimental results with different backbones and input sizes.** Model with EfficientNet-B2 architecture and input size of 768 delivers the best performance.

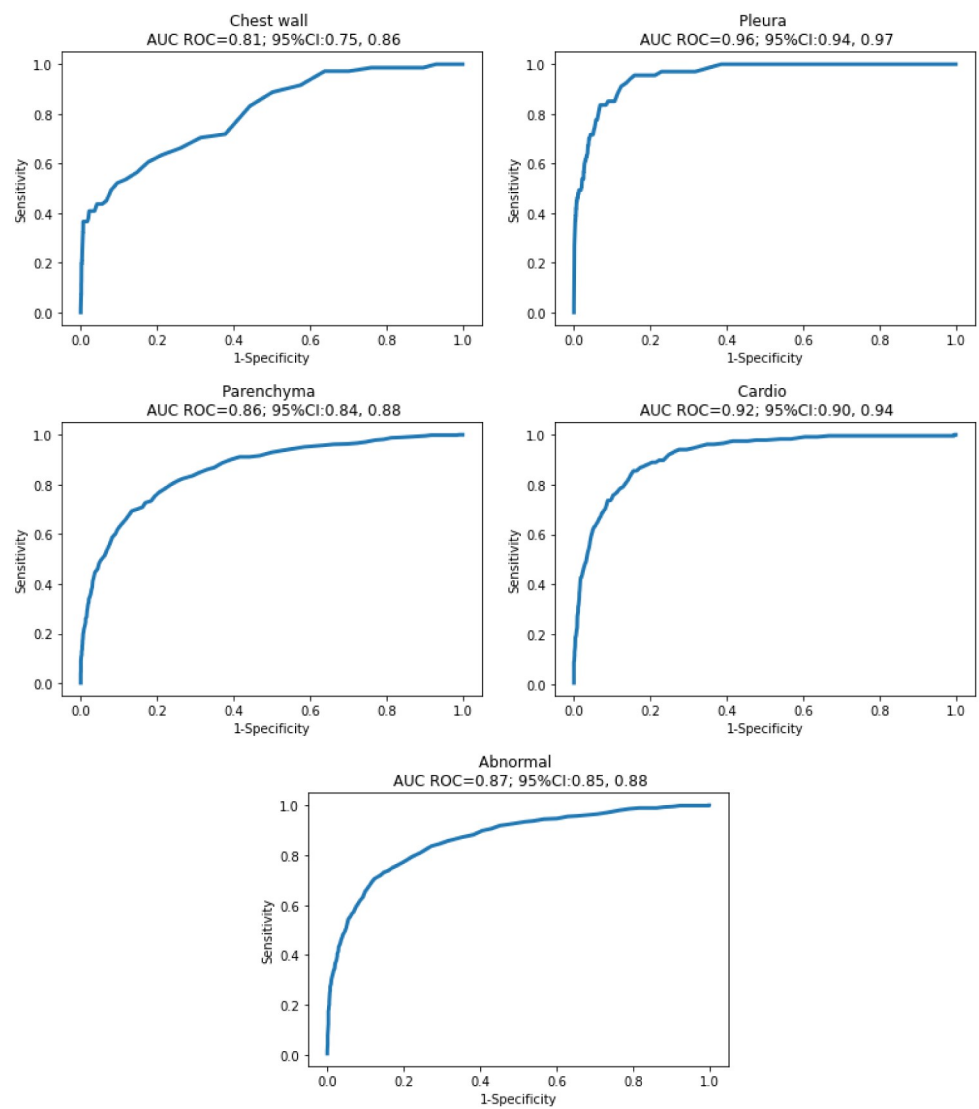
Architecture + Input size	Class	F1 score	AUC	Sensitivity	Specificity
ResNet50 + 768	Bone	0.4348	0.7757	0.3521	0.9935
	Pleura	0.5323	0.9424	0.4925	0.9918
	Parenchyma	0.6274	0.8624	0.6702	0.8706
	Cardio	0.5536	0.9197	0.6043	0.9508
	Abnormal	0.6777	0.8658	0.7512	0.8165
	<b>Average</b>	<b>0.5651</b>	<b>0.8732</b>	<b>0.5741</b>	<b>0.9247</b>
ResNet50 + 1024	Bone	0.3929	0.7879	0.3099	0.9935
	Pleura	0.4593	0.9184	0.4627	0.9874
	Parenchyma	0.6368	0.8599	0.6549	0.8885
	Cardio	0.5521	0.9117	0.6766	0.9342
	Abnormal	0.6856	0.8647	0.684	0.8774
	<b>Average</b>	<b>0.5453</b>	<b>0.8685</b>	<b>0.5576</b>	<b>0.9362</b>
DenseNet121 + 768	Bone	0.2087	0.7097	0.169	0.9891
	Pleura	0.4713	0.9491	0.5522	0.9819
	Parenchyma	0.6003	0.845	0.6656	0.8467
	Cardio	0.4965	0.8909	0.5957	0.9317
	Abnormal	0.6524	0.8446	0.7182	0.8096
	<b>Average</b>	<b>0.4858</b>	<b>0.8479</b>	<b>0.5402</b>	<b>0.9118</b>
DenseNet121 + 1024	Bone	0.1579	0.6892	0.1268	0.9884
	Pleura	0.3515	0.9292	0.6269	0.9557
	Parenchyma	0.5777	0.8298	0.6212	0.8531
	Cardio	0.4624	0.872	0.5362	0.9335
	Abnormal	0.6334	0.8323	0.7252	0.7775
	<b>Average</b>	<b>0.4366</b>	<b>0.8305</b>	<b>0.5272</b>	<b>0.9016</b>
InceptionV3 + 768	Bone	0.2529	0.7198	0.1549	0.9983
	Pleura	0.45	0.9392	0.5373	0.9806
	Parenchyma	0.6015	0.8429	0.6227	0.8757
	Cardio	0.556	0.901	0.5702	0.9591
	Abnormal	0.6606	0.8476	0.6899	0.843
	<b>Average</b>	<b>0.5042</b>	<b>0.8501</b>	<b>0.515</b>	<b>0.9313</b>
InceptionV3 + 1024	Bone	0.3364	0.7454	0.2535	0.9939
	Pleura	0.4379	0.9282	0.5522	0.9778
	Parenchyma	0.6037	0.8386	0.6319	0.8719
	Cardio	0.527	0.8861	0.4979	0.9667
	Abnormal	0.6447	0.8411	0.658	0.849
	<b>Average</b>	<b>0.5099</b>	<b>0.8479</b>	<b>0.5187</b>	<b>0.9319</b>
EfficientNetB2 + 768	Bone	0.4483	0.8035	0.3662	0.9935
	Pleura	0.5085	0.9567	0.4478	0.9928
	Parenchyma	0.6354	0.8602	0.6764	0.8744
	Cardio	0.5597	0.9196	0.6085	0.9519
	Abnormal	0.6970	0.8671	0.6946	0.8825
	<b>Average</b>	<b>0.5698</b>	<b>0.8814</b>	<b>0.5587</b>	<b>0.9390</b>
EfficientNetB2 + 1024	Bone	0.3704	0.7600	0.3521	0.9867
	Pleura	0.5075	0.9352	0.5075	0.9888
	Parenchyma	0.6329	0.8635	0.7178	0.8472
	Cardio	0.5132	0.9129	0.6596	0.9226
	Abnormal	0.6793	0.8595	0.6745	0.8774
	<b>Average</b>	<b>0.5407</b>	<b>0.8662</b>	<b>0.5823</b>	<b>0.9245</b>

<https://doi.org/10.1371/journal.pone.0276545.t005>

**Table 6. Performance of EfficientNet-B2 on five classes.**

Class	Macro F1 score (95%CI)	Average AUC (95%CI)	Average Sensitivity	Average Specificity
Chest wall	0.4483 (0.327, 0.559)	0.8035 (0.748, 0.857)	0.3662	0.9935
Pleura	0.5085 (0.387, 0.617)	0.9567 (0.938, 0.973)	0.4478	0.9928
Parenchyma	0.6354 (0.606, 0.664)	0.8602 (0.843, 0.876)	0.6764	0.8744
Cardio	0.5597 (0.507, 0.608)	0.9196 (0.902, 0.936)	0.6085	0.9519
Abnormal	0.6970 (0.673, 0.722)	0.8671 (0.853, 0.882)	0.6946	0.8825

<https://doi.org/10.1371/journal.pone.0276545.t006>



**Fig 6. Area under the ROC curve.** Pleura class delivered the highest AUC value, at 0.96 (95% CI 0.94, 0.97) whereas chest wall class performed the lowest AUC value, with the figure of 0.81 (95% CI 0.75, 0.85).

<https://doi.org/10.1371/journal.pone.0276545.g006>

**Table 7. The mappings between CheXpert data labels (14 classes) and the proposed set of labels (5 classes).** P and N refer to positive and negative respectively.

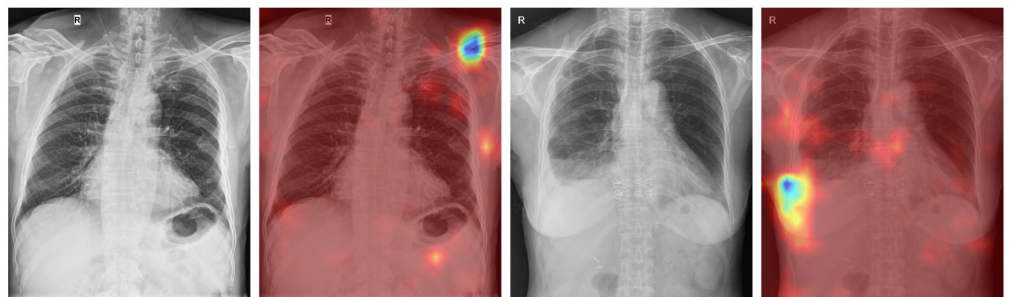
	Chest wall	Pleura	Parenchyma	Cardio	Abnormal
No Finding	N	N	N	N	N
Enlarged Cardiom	N	N	N	P	P
Cardiomegaly	N	N	N	P	P
Lung Lesion	N	N	P	N	P
Lung Opacity	N	N	P	N	P
Edema	N	N	P	N	P
Consolidation	N	N	P	N	P
Pneumonia	N	N	P	N	P
Atelectasis	N	N	P	N	P
Pneumothorax	N	P	N	N	P
Pleural Effusion	N	P	N	N	P
Pleural Other	N	P	N	N	P
Fracture	P	N	N	N	P
Support Devices	N	N	N	N	P

<https://doi.org/10.1371/journal.pone.0276545.t007>

**Table 8. Comparison of coarse and fine classification on CheXpert.**

Architecture	5 classes		12 classes	
	Macro F1 score	F1 score on Abnormal class	Macro F1 score	F1 score on Abnormal class
ResNet50 [40]	0.7109	0.9443	0.4849	0.9444
DenseNet121 [41]	0.7208	0.9519	0.4650	0.9438
InceptionV3 [42]	0.7181	0.9491	0.4846	0.9492
<b>EfficientB2 [43]</b>	<b>0.7429</b>	<b>0.9520</b>	<b>0.5044</b>	<b>0.9450</b>

<https://doi.org/10.1371/journal.pone.0276545.t008>



**Fig 7. Original images and respective Grad-CAMs.** There is a collarbone (nondisplaced fracture) in the first two figures, while the last two ones containing pleural effusion in the pleura. Both of these pathologies were correctly highlighted.

<https://doi.org/10.1371/journal.pone.0276545.g007>

## Conclusion

In current work, we propose a semi-automatic process of building an accurate CXR dataset, which can take advantage of the resources stored in PACS and HIS systems, especially minimizing the intervention of radiologists. We also suggest a coarse classification method based on the location of abnormalities in radiographs, which can address the realistic demand for Vietnamese radiologists and be more efficient than classification based on pathology types. Finally, we demonstrate that building pre-trained models using large CXR datasets can

significantly improve performance compared to using ImageNet datasets. The models fine-tuned from CheXpert pre-trained models with asymmetric loss function achieve significant gains over ImageNet pre-trained models, which we believe will serve as a strong baseline for future research. We also believe that this method will be applied for other languages which have the same characteristic and task requirement.

## Author Contributions

**Conceptualization:** Thang V. Nguyen, Ha Q. Nguyen.

**Data curation:** Thao Nguyen, Thang V. Nguyen.

**Formal analysis:** Thao Nguyen.

**Methodology:** Thao Nguyen, Tam M. Vo, Thang V. Nguyen, Hieu H. Pham, Ha Q. Nguyen.

**Supervision:** Ha Q. Nguyen.

**Validation:** Thao Nguyen, Tam M. Vo, Thang V. Nguyen, Ha Q. Nguyen.

**Visualization:** Tam M. Vo.

**Writing – original draft:** Tam M. Vo, Hieu H. Pham, Ha Q. Nguyen.

**Writing – review & editing:** Hieu H. Pham, Ha Q. Nguyen.

## References

1. Delrue L, Gosselin R, Ilsen B, Van Landeghem A, de Mey J, Duyck P. Difficulties in the interpretation of chest radiography. In *Comparative interpretation of CT and standard radiography of the chest 2011* (pp. 27–49). Springer, Berlin, Heidelberg.
2. American College of Radiology. ACR–SPR–STR practice parameter for the performance of chest radiography 2011; Available at: <https://www.acr.org/-/media/ACR/Files/Practice-Parameters/ChestRad.pdf>. Accessed August 22, 2021.
3. Tran, Thanh T and Pham, Hieu H and Nguyen, Thang V and Le, Tung T and Nguyen, Hieu T and Nguyen, Ha Q. Learning to automatically diagnose multiple diseases in pediatric chest radiographs using deep convolutional neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (3314–3323), 2021.
4. Pham, Hieu H and Nguyen, Ha Q and Nguyen, Hieu T and Le, Linh T and Khanh, Lam. An Accurate and Explainable Deep Learning System Improves Interobserver Agreement in the Interpretation of Chest Radiograph. *arXiv:2208.03545*, 2022.
5. Le, Khiem H and Tran, Tuan V and Pham, Hieu H and Nguyen, Hieu T and Le, Tung T and Nguyen, Ha Q. Learning from Multiple Expert Annotators for Enhancing Anomaly Detection in Medical Image Analysis. *arXiv preprint arXiv:2203.10611*, 2022.
6. Rajpurkar, Pranav and Joshi, Anirudh and Pareek, Anuj and Chen, Phil and Kiani, Amirhossein and Irvin, Jeremy et al. CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. *arXiv preprint arXiv:2002.11379*, 2020.
7. Wang, Hongyu and Xia, Yong. Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. *arXiv preprint arXiv:1807.03058*, 2018.
8. Tang Yu-Xing and Tang You-Bao and Peng Yifan and Yan Ke and Bagheri Mohammadhadi and Redd Bernadette A et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digital Medicine (Nature Publishing Group)*, 2018.
9. Liang Gaobo and Zheng Lixin. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*, (pp. 104–964), 2020.
10. Nguyen, Hieu T and Pham, Hieu H and Nguyen, Nghia T and Nguyen, Ha Q and Huynh, Thang Q and Dao, Minh et al. VinDr-SpineXR: A deep learning framework for spinal lesions detection and classification from radiographs. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (pp. 291–301), 2021.
11. Pham Hieu H and Do Dung V and Nguyen Ha Q. Dicom imaging router: An open deep learning framework for classification of body parts from dicom x-ray scans. *medRxiv*, 2021.

12. Nguyen, Huyen TX and Tran, Sam B and Nguyen, Dung B and Pham, Hieu H and Nguyen, Ha Q. A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms. arXiv preprint arXiv:2112.04490, 2021.
13. Shen Li and Margolies Laurie R and Rothstein Joseph H and Fluder Eugene and McBride Russell and Sieh Weiva. Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports* (Nature Publishing Group, (pp. 1–12) 2019. <https://doi.org/10.1038/s41598-019-48995-4> PMID: 31467326
14. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence 2019 Jul 17* (Vol. 33, No. 01, pp. 590–597).
15. Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng Y, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042. 2019 Jan 21.
16. Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Med Image Anal.* 2020; 66:101797. <https://doi.org/10.1016/j.media.2020.101797> PMID: 32877839
17. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2017* (pp. 2097–2106).
18. Nguyen, Hoang C and Le, Tung T and Pham, Hieu H and Nguyen, Ha Q. VinDr-RibCXR: A Benchmark Dataset for Automatic Segmentation and Labeling of Individual Ribs on Chest X-rays. arXiv preprint arXiv:2107.01327, 2021.
19. Nguyen HQ, Lam K, Le LT, Pham HH, Tran DQ, Nguyen DB, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. arXiv preprint arXiv:2012.15029. 2020 Dec 30.
20. Nguyen, H. Q., Pham, H. H., Tuan Linh, L., Dao, M., Khanh, L. VinDr-CXR: An open dataset of chest X-rays with radiologist annotations (version 1.0.0) *PhysioNet*, <https://doi.org/10.13026/3akn-b287>, 2021.
21. Nguyen, Ngoc H and Pham, Hieu H and Tran, Thanh T and Nguyen, Tuan NM and Nguyen, Ha Q. VinDr-PCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. arXiv preprint arXiv:2203.10612, 2022.
22. Pham, H. H., Tran, T. T., Nguyen, H. Q. VinDr-PCXR: An open, large-scale pediatric chest X-ray dataset for interpretation of common thoracic diseases. *PhysioNet* (version 1.0.0), <https://doi.org/10.13026/k8qc-na36>, 2022
23. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings.* 2018; 2018:188. PMID: 29888070
24. Filice RW, Stein A, Wu CC, Arteaga VA, Borstelmann S, et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. *Journal of digital imaging.* 2020 Apr; 33(2):490–6. <https://doi.org/10.1007/s10278-019-00299-9> PMID: 31768897
25. Jain S, Agrawal A, Saporta A, Truong SQ, Bui T, Chambon P, et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *Conference on Neural Information Processing Systems (NeurIPS 2021)*
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
27. McDermott MB, Hsu TM, Weng WH, Ghassemi M, Szolovits P. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference 2020 Sep 18* (pp. 913–927). PMLR.
28. Jain S, Smit A, Truong SQ, Nguyen CD, Huynh MT, Jain M, et al. VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning 2021 Apr 8* (pp. 105–115).
29. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:2004.09167. 2020 Apr 20.
30. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. *Natural Language Engineering.* 1995 Mar; 1(1):83–108. <https://doi.org/10.1017/S1351324900000061>
31. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and



- applications. *Journal of the American Medical Informatics Association*. 2010 Sep 1; 17(5):507–13. <https://doi.org/10.1136/jamia.2009.001560> PMID: 20819853
32. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Illcus S, Chute C, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 33; 2019. p. 590–597.
  33. Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, et al. CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. 2020. arXiv:2002.11379 [eess.IV].
  34. Pham HH, Le TT, Tran DQ, Ngo DT, Nguyen HQ. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. arXiv preprint arXiv:191106475. 2020.
  35. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng Cy, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*. 2019; 6(1):317. <https://doi.org/10.1038/s41597-019-0322-0> PMID: 31831740
  36. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*. 2018; 15(11):1–17. <https://doi.org/10.1371/journal.pmed.1002686> PMID: 30457988
  37. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, et al. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*. 2020; 294(2):421–431. <https://doi.org/10.1148/radiol.2019191293> PMID: 31793848
  38. Nguyen NH, Nguyen HQ, Nguyen NT, Nguyen TV, Pham HH, Nguyen TN. A clinical validation of VinDr-CXR, an AI system for detecting abnormal chest radiographs. arXiv preprint arXiv:2104.02256. 2021 Apr 6.
  39. Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning 2021* Apr 8 (pp. 116–124).
  40. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2016* (pp. 770–778).
  41. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2017* (pp. 4700–4708).
  42. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2016* (pp. 2818–2826).
  43. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning 2019* May 24 (pp. 6105–6114). PMLR.
  44. Ben-Baruch E, Ridnik T, Zamir N, Noy A, Friedman I, Protter M, et al. Asymmetric loss for multi-label classification. arXiv preprint arXiv:2009.14119. 2020 Sep 29.
  45. Wu T, Huang Q, Liu Z, Wang Y, Lin D. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision 2020* Aug 23 (pp. 162–178). Springer, Cham.
  46. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. No. 57 in *Monographs on Statistics and Applied Probability*. Boca Raton, Florida, USA: Chapman & Hall/CRC; 1993.
  47. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision 2017* (pp. 618–626).