



OPEN

DATA DESCRIPTOR

VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations

Ha Q. Nguyen^{1,2,9}, Khanh Lam^{3,9}, Linh T. Le^{4,9}, Hieu H. Pham^{1,2,5,6}✉, Dat Q. Tran¹, Dung B. Nguyen¹, Dung D. Le^{3,10}, Chi M. Pham^{3,10}, Hang T. T. Tong^{3,10}, Diep H. Dinh^{3,10}, Cuong D. Do^{3,10}, Luu T. Doan^{4,10}, Cuong N. Nguyen^{4,10}, Binh T. Nguyen^{4,10}, Que V. Nguyen^{4,10}, Au D. Hoang^{4,10}, Hien N. Phan^{4,10}, Anh T. Nguyen^{4,10}, Phuong H. Ho^{7,10}, Dat T. Ngo², Nghia T. Nguyen², Nhan T. Nguyen², Minh Dao¹ & Van Vu^{1,8}

Most of the existing chest X-ray datasets include labels from a list of findings without specifying their locations on the radiographs. This limits the development of machine learning algorithms for the detection and localization of chest abnormalities. In this work, we describe a dataset of more than 100,000 chest X-ray scans that were retrospectively collected from two major hospitals in Vietnam. Out of this raw data, we release 18,000 images that were manually annotated by a total of 17 experienced radiologists with 22 local labels of rectangles surrounding abnormalities and 6 global labels of suspected diseases. The released dataset is divided into a training set of 15,000 and a test set of 3,000. Each scan in the training set was independently labeled by 3 radiologists, while each scan in the test set was labeled by the consensus of 5 radiologists. We designed and built a labeling platform for DICOM images to facilitate these annotation procedures. All images are made publicly available in DICOM format along with the labels of both the training set and the test set.

Background & Summary

Computer-aided diagnosis (CAD) systems for chest radiographs (also referred to as Chest X-ray or CXR) have recently achieved great success thanks to the availability of large labeled datasets and the recent advances of high-performance supervised learning algorithms^{1–5}. Leveraging deep convolutional neural networks (CNN)⁶, these systems can reach the expert-level performance in classifying common lung diseases and related findings. Training a CNN heavily relies on high quality datasets of annotated images. However, it is costly and time-consuming to build such datasets due to several constraints: (1) medical data are hard to retrieve from hospitals or medical centers; (2) manual annotation by physicians is expensive; (3) the annotation of medical images requires a consensus of several expert readers to overcome human bias⁷; and (4) it lacks an efficient labeling framework to manage and annotate large-scale medical datasets.

Notable public datasets of CXR include ChestX-ray8, ChestX-ray14⁸, Padchest⁹, CheXpert², and MIMIC-CXR¹⁰. ChestX-ray14, an extended version of ChestX-ray8, was released by the US National Institutes of Health (NIH), containing over 112,000 CXR scans from more than 30,000 patients. Without being manually annotated, this dataset poses significant issues related to the quality of its labels¹¹. Padchest consists of more than 160,000 CXR images, 27% of which were hand-labeled by radiologists with 174 different findings and 19 diagnoses. The rest of the dataset were labeled using a Natural Language Processing (NLP) tool. Recently released CheXpert provides more than 200,000 CXRs of 65,240 patients, which were labeled for the presence of

¹Vingroup Big Data Institute, Hanoi, Vietnam. ²Smart Health Center, VinBigData JSC, Hanoi, Vietnam. ³Hospital 108, Department of Radiology, Hanoi, Vietnam. ⁴Hanoi Medical University Hospital, Department of Radiology, Hanoi, Vietnam. ⁵College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam. ⁶VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam. ⁷Tam Anh General Hospital, Department of Radiology, Ho Chi Minh City, Vietnam. ⁸Yale University, Department of Mathematics, New Heaven, CT, 06511, USA. ⁹These authors contributed equally: Ha Q. Nguyen, Khanh Lam, Linh T. Le. ¹⁰These authors contributed equally: Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho. ✉e-mail: hieu.ph@vinuni.edu.vn

Dataset	Release year	# findings	# samples	Image-level labels	Local labels
JSRT ¹²	2000	1	247 ^(*)	Available	Available
MC ¹⁴	2014	1	138 ^(*)	Available	N/A
SH ¹⁴	2014	1	662 ^(*)	Available	N/A
Indiana ¹³	2016	10	8,121 ^(*)	Available	N/A
ChestX-ray8 ⁸	2017	8	108,948 ^(*)	Available	Available [†]
ChestX-ray14 ⁸	2017	14	112,120 ^(*)	Available	N/A
CheXpert ²	2019	14	224,316 ^(*)	Available	N/A
Padchest ⁹	2019	193	160,868 ^(*)	Available	N/A ^{††}
MIMIC-CXR ¹⁰	2019	14	377,110 ^(*)	Available	N/A
VinDr-CXR (ours)	2020	28	18,000 ^(*)	Available	Available

Table 1. An overview of existing public datasets for CXR interpretation. ● Labeled by an NLP algorithm. (*) Labeled by radiologists. (†) Moderate-size datasets that are not applicable for training deep learning models. (††) A portion of the dataset (983 images) is provided with hand-labeled bounding boxes. (††) 27% of the dataset was manually annotated with encoded anatomical regions of the findings.

14 observations using an automated rule-based labeler that extracts keywords from medical reports. Adopting the same labeling mechanism, MIMIC-CXR contains 377,110 images in DICOM format along with free-text radiology reports. Table 1 provides a summary of the aforementioned datasets together with other ones of moderate sizes, including JSRT¹², Indiana¹³, MC¹⁴, and SH¹⁴.

Most existing CXR datasets depend on automated rule-based labelers that either use keyword matching (e.g. CheXpert² and NIH labelers⁸) or an NLP model to extract disease labels from free-text radiology reports. These tools can produce labels on a large scale but, at the same time, introduce a high rate of inconsistency, uncertainty, and errors^{11,15}. These noisy labels may lead to the deviation of deep learning-based algorithms from reported performances when evaluated in a real-world setting¹⁶. Furthermore, the report-based approaches only associate a CXR image with one or several labels in a predefined list of findings and diagnoses without identifying their locations. There are a few CXR datasets that include annotated locations of abnormalities but they are either too small for training deep learning models (JSRT) or not detailed enough (PadChest). The interpretation of a CXR is not all about image-level classification; it is even more important, from the perspective of a radiologist, to localize the abnormalities on the image. This partly explains why the applications of CAD systems for CXR in clinical practice are still very limited.

In an effort to provide a large CXR dataset with high-quality labels for the research community, we have built the VinDr-CXR dataset from more than 100,000 raw images in DICOM format that were retrospectively collected from the Hospital 108 (H108) and the Hanoi Medical University Hospital (HMHU), two of the largest hospitals in Vietnam. The published dataset consists of 18,000 postero-anterior (PA) view CXR scans that come with both the localization of critical findings and the classification of common thoracic diseases. These images were annotated by a group of 17 radiologists with at least 8 years of experience for the presence of 22 critical findings (local labels) and 6 diagnoses (global labels); each finding is localized with a bounding box. The local and global labels correspond to the “Findings” and “Impressions” sections, respectively, of a standard radiology report. We divide the dataset into two parts: the training set of 15,000 scans and the test set of 3,000 scans. Each image in the training set was independently labeled by 3 radiologists, while the annotation of each image in the test set was even more carefully treated and obtained from the consensus of 5 radiologists. The labeling process was performed via an in-house system called VinDr Lab¹⁷, which was built on top of a Picture Archiving and Communication System (PACS). All DICOM images and the labels of both the training set and the test set are released. A slightly modified version of this dataset was used to organize the VinBigData Chest Xray Abnormalities Detection Challenge on the Kaggle platform (<https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/>).

VinDr-CXR, to the best of our knowledge, is currently the largest public CXR dataset with radiologist-generated annotations in both training and test sets. We believe the dataset will accelerate the development and evaluation of new machine learning models for both localization and classification of thoracic lesions and diseases on CXR scans.

Methods

The building of VinDr-CXR dataset, as visualized in Fig. 1, is divided into three main steps: (1) data collection, (2) data filtering, and (3) data labeling. Between 2018 and 2020, we retrospectively collected more than 100,000 CXRs in DICOM format from local PACS servers of two hospitals in Vietnam, HMHU and H108. Imaging data were acquired from a wide diversity of scanners from well-known medical equipment manufacturers, including Phillips, GE, Fujifilm, Siemens, Toshiba, Canon, Samsung, and Carestream. The ethical clearance of this study was approved by the Institutional Review Boards (IRBs) of the HMHU and H108 before the study started. The need for obtaining informed patient consent was waived because this retrospective study did not impact clinical care or workflow at these two hospitals and all patient-identifiable information in the data has been removed.

Data de-identification. To protect patient’s privacy¹⁸, all personally identifiable information associated with the images has been removed or replaced with random values. Specifically, we ran a Python script that removes

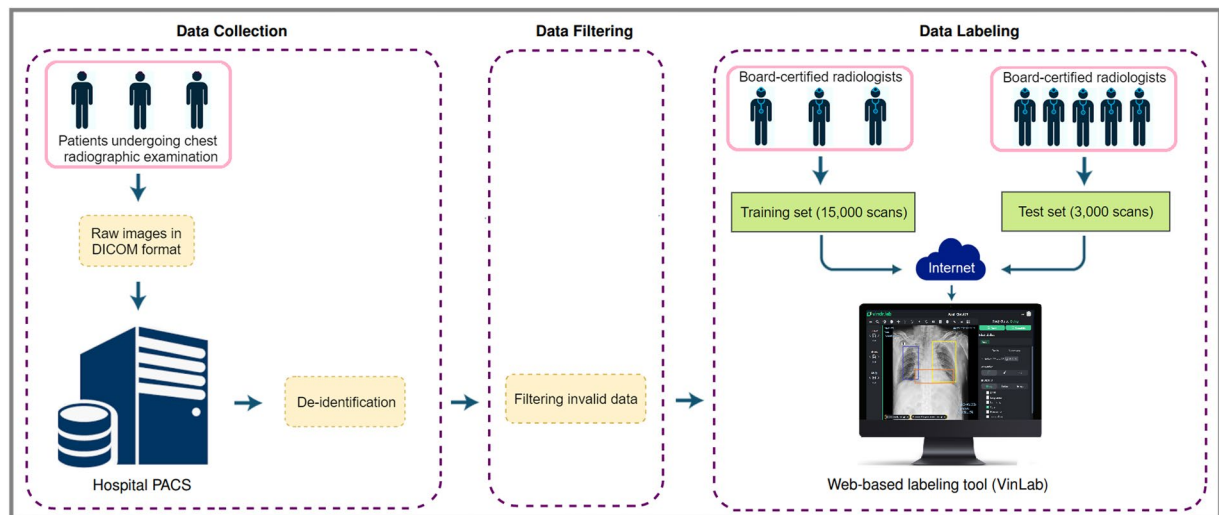


Fig. 1 The flow of creating VinDr-CXR dataset: (1) raw images in DICOM format were collected retrospectively from the hospital's PACS and got de-identified to protect patient's privacy; (2) invalid files, such as images of other modalities, other body parts, low quality, or incorrect orientation, were automatically filtered out by a CNN-based classifier; (3) A web-based labeling tool, VinDr Lab, was developed to store, manage, and remotely annotate DICOM data: each image in the training set of 15,000 images was independently labeled by a group of 3 radiologists and each image in the test set of 3,000 images was labeled by the consensus of 5 radiologists.

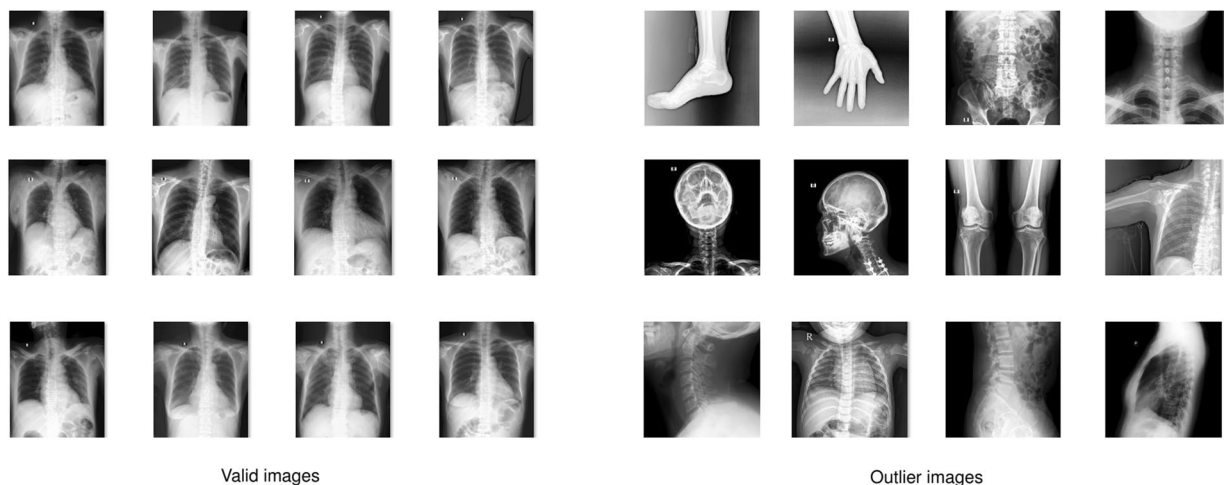


Fig. 2 Examples of valid (**left**) and invalid (**right**) CXR scans. A CNN-based classifier was trained and used to automatically filter outliers; only valid PA-view CXRs of adults were retained for labeling.

all DICOM tags of protected health information (PHI)¹⁹ such as: patient's name, patient's date of birth, patient ID, or acquisition time and date, etc. We only retained a limited number of DICOM attributes that are necessary for processing raw images. The entire list of retained attributes is shown in Table 1 (supplementary materials). Next, a simple algorithm was implemented to automatically remove textual information appearing on the image data (i.e. pixel annotations that could include patient's identifiable information). The resulting images were then manually verified to make sure all texts were removed before they were digitally sent out of the hospitals' systems.

Data filtering. The collected raw data was mostly of *adult PA-view CXRs*, but also included a significant amount of outliers such as images of body parts other than chest (due to mismatched DICOM tags), pediatric scans, low-quality images, or lateral CXRs. Examples of these images are shown in Fig. 2. All outliers were automatically excluded from the dataset using a binary classifier, which is a light-weight convolutional neural network (CNN). The training procedure of this classifier is out of the scope of this paper.

Data labeling. The VinDr-CXR dataset was labeled for a total of 28 findings and diagnoses in adult cases: (1) Aortic enlargement, (2) Atelectasis, (3) Cardiomegaly, (4) Calcification, (5) Clavicle fracture, (6) Consolidation, (7) Edema, (8) Emphysema, (9) Enlarged PA, (10) Interstitial lung disease (ILD), (11) Infiltration, (12) Lung

	Characteristics	Training set	Test set
Collection statistics	Years	2018 to 2020	2018 to 2020
	Number of scans	15,000	3,000
	Number of human annotators per scan	3	5
	Image size (pixel × pixel, median)	2788 × 2446	2748 × 2394
	Age (years, median)*	43.77	31.80
	Male (%)*	52.21	55.90
	Female (%)*	47.79	44.10
	Data size (GB)	161	31.3
Local labels	1. Aortic enlargement (%)	2348 (15.65%)	220 (7.33%)
	2. Atelectasis (%)	62 (0.41%)	86 (2.87%)
	3. Cardiomegaly (%)	1817 (12.11%)	309 (10.30%)
	4. Calcification (%)	177 (1.18%)	194 (6.47%)
	5. Clavicle fracture (%)	1 (0.01%)	2 (0.07%)
	6. Consolidation (%)	121 (0.81%)	96 (3.20%)
	7. Edema (%)	1 (0.01%)	0 (0%)
	8. Emphysema (%)	14 (0.09%)	3 (0.1%)
	9. Enlarged PA (%)	21 (0.14%)	8 (0.27%)
	10. Interstitial lung disease (ILD) (%)	152 (1.01%)	221 (7.37%)
	11. Infiltration (%)	245 (1.63%)	58 (1.93%)
	12. Lung cavity (%)	21 (0.14%)	9 (0.30%)
	13. Lung cyst (%)	4 (0.03%)	2 (0.07%)
	14. Lung opacity (%)	547 (3.65%)	84 (2.80%)
	15. Mediastinal shift (%)	85 (0.57%)	20 (0.67%)
	16. Nodule/Mass (%)	410 (2.73%)	176 (5.87%)
	17. Pulmonary fibrosis (%)	1017 (6.78%)	217 (7.23%)
	18. Pneumothorax (%)	58 (0.39%)	18 (0.60%)
	19. Pleural thickening (%)	882 (5.88%)	169 (5.63%)
	20. Pleural effusion (%)	634 (4.23%)	111 (3.70%)
	21. Rib fracture (%)	41 (0.27%)	11 (0.37%)
	22. Other lesion (%)	363 (2.42%)	94 (3.13%)
Global labels	23. Lung tumor (%)	132 (0.88%)	80 (2.67%)
	24. Pneumonia (%)	469 (3.13%)	246 (8.20%)
	25. Tuberculosis (%)	479 (3.19%)	164 (5.47%)
	26. Other diseases (%)	4002 (26.68%)	657 (21.90%)
	27. COPD (%)	7 (0.05%)	2 (0.07%)
	28. No finding (%)	10606 (70.71%)	2051 (68.37%)

Table 2. Dataset characteristics. Note: the numbers of positive labels were reported based on the majority vote of the participating radiologists. (*) The calculations were only based on the CXR scans where patient's sex and age were known.

cavity, (13) Lung cyst, (14) Lung opacity, (15) Mediastinal shift, (16) Nodule/Mass, (17) Pulmonary fibrosis, (18) Pneumothorax, (19) Pleural thickening, (20) Pleural effusion, (21) Rib fracture, (22) Other lesion, (23) Lung tumor, (24) Pneumonia, (25) Tuberculosis, (26) Other diseases, (27) Chronic obstructive pulmonary disease (COPD), and (28) No finding. These labels were divided into 2 categories: local labels (1–22) and global labels (23–28). The local labels should be marked with bounding boxes that localize the findings, while the global labels should reflect the diagnostic impression of the radiologist. The definition of each label is detailed in Table 2 (supplementary materials). This list of labels was suggested by a committee of the most experienced radiologists from the two hospitals. The selection of these labels took into account two factors: first, they are prevalent and second, they can be differentiated on CXRs. Figure 3 illustrates several samples with both local and global labels annotated by radiologists.

To facilitate the labeling process, we designed and built a web-based framework called VinDr Lab and had a team of 17 experienced radiologists remotely annotate the data. All the radiologists participating in the labeling process were certified in diagnostic radiology and received healthcare profession certificates from the Vietnamese Ministry of Health. A set of 18,000 CXRs were randomly chosen from the filtered data, of which 15,000 scans (normal: 10,606 studies, abnormal: 4394 studies) serve as the training set and the remaining 3,000 (normal: 2052 studies, abnormal: 948 studies) form the test set. Each sample in the training set was assigned to 3 radiologists for annotating in a blind fashion. Additionally, all of the participating radiologists were blinded to relevant clinical information. For the test set, 5 radiologists were involved in a two-stage labeling process.

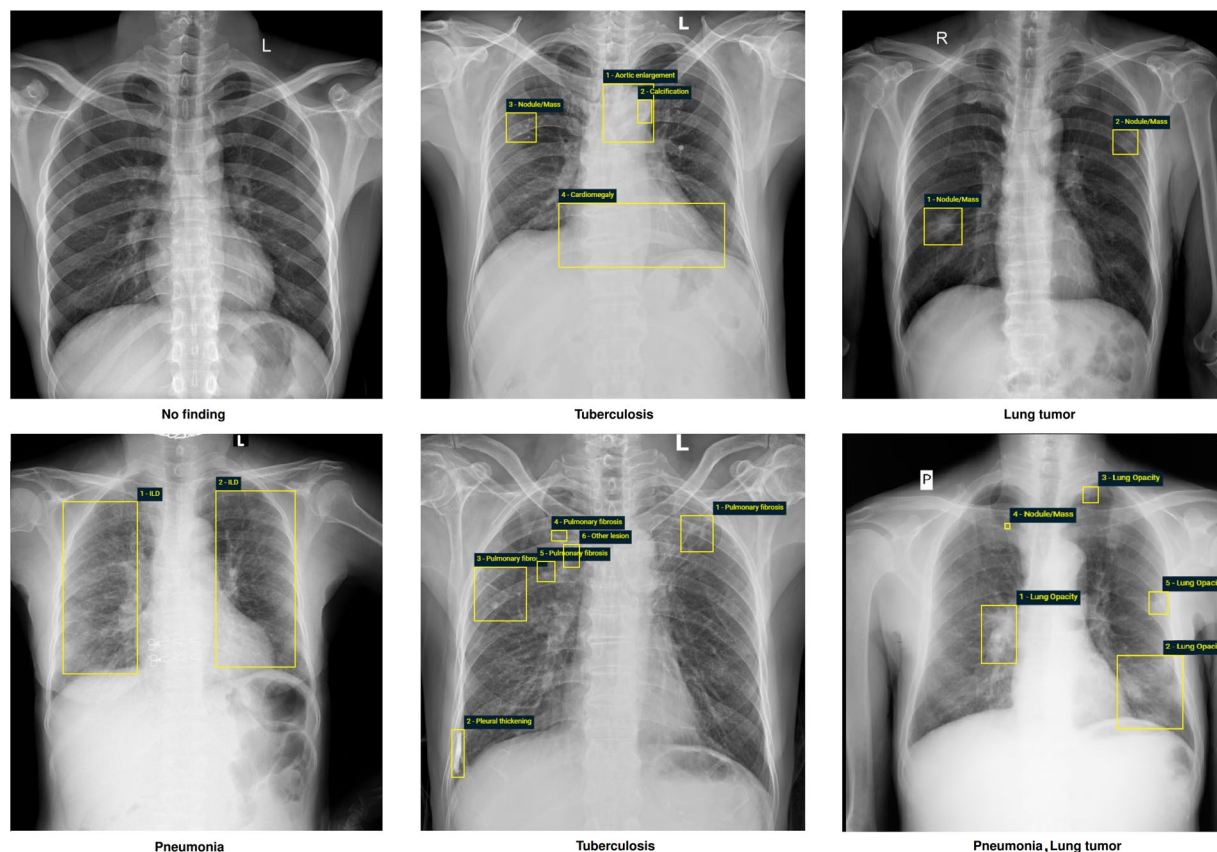


Fig. 3 Examples of CXRs with radiologist’s annotations. Abnormal findings (local labels) marked by radiologists are plotted on the original images for visualization purpose. The global labels are in bold and listed at the bottom of each example. Better viewed on a computer and zoomed in for details.

During the first stage, each image was independently annotated by 3 radiologists. In the second stage, 2 other radiologists, who have a higher level of experience, reviewed the annotations of the 3 previous annotators and communicated with each other in order to decide the final labels. The disagreements among initial annotators, as shown in Fig. 3 (supplementary materials), were carefully discussed and resolved by the 2 reviewers. Finally, the consensus of their opinions will serve as reference ground-truth.

Once the labeling was completed, the labels of 18,000 CXRs were exported in JavaScript Object Notation (JSON) format. We then parsed their contents and organized the annotations in the form of a single comma-separated values (CSV) file. As a result, we provided a single CSV file that contains labels, bounding box coordinates, and their corresponding image IDs. For the training set, each sample comes with the annotations of three different radiologists. For the test set, we only provide the consensus labels of the five radiologists. The data characteristics, including patient demographic and the prevalence of each finding or pathology, are summarized in Table 2. The distribution of all labels in the training set is drawn in Fig. 4. We have released all images together with the labels of the training set and the test set.

Data Records

The VinDr-CXR dataset has been submitted to PhysioNet for public download²⁰. We provide all imaging data and the corresponding ground truth labels for both the training and test sets. The images were organized into two folders, one for training and the other one for testing. Each image has a unique, anonymous identifier which was encoded from the value of the SOP Instance UID provided by the DICOM tag (0008,0018). The encoding process was supported by the Python hashlib module (see [sec:code]Code Availability). The radiologists’ local annotations of the training set were provided in a CSV file, annotations_train.csv. Each row of the table represents a bounding box with the following attributes: image ID (image_id), radiologist ID (rad_id), label’s name (class_name), and bounding box coordinates (x_min, y_min, x_max, y_max). Here, rad_id encodes the identities of the 17 radiologists, (x_min, y_min) are the coordinates of the box’s upper left corner, and (x_max, y_max) are the coordinates of the lower-right corner. Meanwhile, the image-level labels of the training set were stored in different CSV file, image_labels_train.csv, with the following fields: Image ID (image_id), radiologist ID (rad_ID), and labels (labels) for both the findings and diagnoses. Specifically, each image ID goes with vector of multiple labels corresponding to different pathologies, in which positive ones were encoded with “1” and negative ones were encoded with “0”. Similarly, the bounding-box annotations and the image-level labels of the test set were recorded in annotations_test.csv and image_labels_test.csv, respectively. The only difference is that each row in the CSV files of the test set was not associated with a radiologist ID.

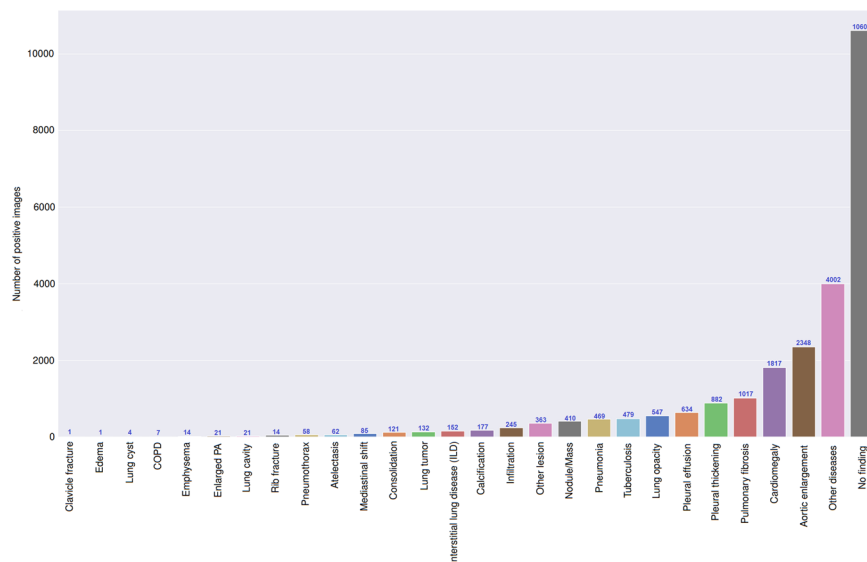


Fig. 4 Distribution of findings and pathologies on the training set of VinDr-CXR.

Technical Validation

The data de-identification was controlled. In particular, all DICOM meta-data was parsed and manually reviewed to ensure that all individually identifiable health information of the patients has been removed to meet the U.S. HIPAA (<https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>), the European GDPR (<https://gdpr-info.eu/>), as well as the local privacy laws. Pixel values of all CXR scans were also carefully examined. All images were manually reviewed case-by-case by a team of 10 human readers. During this review process, a small number of images containing private textual information that had not been removed by our algorithm was excluded from the dataset. The manual review process also helped identify and discard outlier samples that the CNN-based classifier was not able to detect. To control the quality of the labeling process, we developed a set of rules underlying VinDr Lab for automatic verification of radiologist-generated labels. These rules prevent annotators from mechanical mistakes like forgetting to choose global labels or marking lesions on the image while choosing “No finding” as the global label. To ensure the complete blindness among annotators, the images were randomly shuffled before being assigned to each of them.

Usage Notes

To download the dataset, users are required to accept a Data Usage Agreement (DUA) called PhysioNet Credentialed Health Data License 1.5.0 (<https://physionet.org/content/vindr-cxr/view-license/1.0.0/>). By accepting the DUA, users agree that they will not share the data and that the dataset can be used for scientific research and educational purposes only and will not attempt to re-identify any patients, institutions or hospitals. For any publication that explores this resource, the authors must cite this original paper. We also encourage such authors to release their code and models, which will help the community to reproduce experiments and to boost the research in the field of medical imaging.

Code availability

The code used for loading and processing DICOM images is based on the following open-source repositories: Python 3.7.0 (<https://www.python.org/>); Pydicom 1.2.0 (<https://pydicom.github.io/>); OpenCV-Python 4.2.0.34 (<https://pypi.org/project/opencv-python/>); and Python hashlib (<https://docs.python.org/3/library/hashlib.html>). The code for data de-identification and outlier detection was made publicly available at <https://github.com/vinbigdata-medical/vindr-cxr>.

Received: 4 August 2021; Accepted: 23 June 2022;

Published online: 20 July 2022

References

- Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine* **15**, e1002686, <https://doi.org/10.1371/journal.pmed.1002686> (2018).
- Irvin, J. *et al.* CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 590–597 (2019).
- Majkowska, A. *et al.* Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* **294**, 421–431, <https://doi.org/10.1148/radiol.2019191293> (2020).
- Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine* **3**, 1–8, <https://doi.org/10.1038/s41746-020-0273-z> (2020).
- Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T. & Nguyen, H. Q. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing* **437**, 186–194 (2021).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **512**, 436–444, <https://doi.org/10.1038/nature14539> (2015).

7. Razzak, M. I., Naz, S. & Zaib, A. Deep learning for medical image processing: Overview, challenges and the future. In *Classification in BioApps*, 323–350, https://doi.org/10.1007/978-3-319-65981-7_12 (Springer, 2018).
8. Wang, X. *et al.* 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 (CVPR)*, 2097–2106, <https://doi.org/10.1109/CVPR.2017.369> (2017).
9. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* **66**, 101797 (2020).
10. Johnson, A. E. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317, <https://doi.org/10.1038/s41597-019-0322-0> (2019).
11. Oakden-Rayner, L. Exploring the ChestXray14 dataset: problems. <https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/> (2017). (Online; accessed 04 May 2020).
12. Shiraishi, J. *et al.* Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology* **174**, 71–74, <https://doi.org/10.2214/ajr.174.1.1740071> (2000).
13. Demner-Fushman, D. *et al.* Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **23**, 304–310, <https://doi.org/10.1093/jamia/ocv080> (2016).
14. Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery* **4**, 475–477 <https://doi.org/10.3978/j.issn.2223-4292.2014.11.20> (2014).
15. Oakden-Rayner, L. Exploring large-scale public medical image datasets. *Academic Radiology* **27**, 106–112, <https://doi.org/10.1016/j.acra.2019.10.006>. Special Issue: Artificial Intelligence (2020).
16. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* **368**, <https://doi.org/10.1136/bmj.m689> (2020).
17. Nguyen, N. T. *et al.* VinDr Lab: A data platform for medical AI. <https://github.com/vinbigdata-medical/vindr-lab> (2021).
18. Vietnamese National Assembly. Regulation 40/2009/QH12 (Law on Medical Examination and Treatment). <http://vbpl.vn/hanoi/Pages/vbpqen-toanvan.aspx?ItemID=10482>. (Online; accessed 11 December 2020) (2009).
19. Isola, S. & Al Khalili, Y. Protected Health Information (PHI). <https://www.ncbi.nlm.nih.gov/books/NBK553131/> (2019).
20. Nguyen, H. Q., Pham, H. H., Le, L. T., Dao, M. & Lam, K. VinDr-CXR: An open dataset of chest X-rays with radiologist annotations. *PhysioNet* <https://doi.org/10.13026/3akn-b287> (2021).

Acknowledgements

The authors would like to acknowledge the Hanoi Medical University Hospital and the Hospital 108 for providing us access to their image databases and for agreeing to make the VinDr-CXR dataset publicly available. We are especially thankful to all of our collaborators, including radiologists, physicians, and technicians, who participated in the data collection and labeling process.

Author contributions

H.Q.N., K.L. and L.L. designed the study; H.Q.N., Nghia T. Nguyen, M.D. and V.V. designed the labeling framework; H.H.P. and D.B.N. performed the data de-identification; H.H.P. developed the algorithm for outlier filtering; D.T., D.B.N., D.T.N. and Nhan T. Nguyen conducted the data acquisition and analysis; K.L., L.L., D.L., C.P., H.T., D.D., C.D., L.D., C.N., B.N., Q.N., A.H., H.N.P., A.N. and P.H. annotated data and made comments to improve the labeling tools; H.Q.N. and H.H.P. wrote the paper; all authors reviewed the manuscript.

Competing interests

This work was funded by the Vingroup JSC. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-022-01498-w>.

Correspondence and requests for materials should be addressed to H.H.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022