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

Ha Q. Nguyen[1,2,†], Khanh Lam[3,†], Linh T. Le[4,†], Hieu H. Pham[1,2,*], Dat Q. Tran[1], Dung B. Nguyen[1], Dung D. Le[3,‡], Chi M. Pham[3,‡], Hang T. T. Tong[3,‡], Diep H. Dinh[3,‡], Cuong D. Do[3,‡], Luu T. Doan[4,‡], Cuong N. Nguyen[4,‡], Binh T. Nguyen[4,‡], Que V. Nguyen[4,‡], Au D. Hoang[4,‡], Hien N. Phan[4,‡], Anh T. Nguyen[4,‡], Phuong H. Ho[5,‡], Dat T. Ngo[1], Nghia T. Nguyen[1], Nhan T. Nguyen[1], Minh Dao[1], and Van Vu[1,6]

[1]Vingroup Big Data Institute (VinBigdata), Hanoi, Vietnam
[2]VinUniversity, Hanoi, Vietnam
[3]Hospital 108, Department of Radiology, Hanoi, Vietnam
[4]Hanoi Medical University Hospital, Department of Radiology, Hanoi, Vietnam
[5]Tam Anh General Hospital, Department of Radiology, Ho Chi Minh City, Vietnam
[6]Yale University, Department of Mathematics, New Heaven, CT 06511, U.S.A.
[*]corresponding author: Hieu H. Pham (hieuph4@vinbigdata.org)
[†]these authors contributed equally to this work
[‡]these authors contributed equally to this work

ABSTRACT

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 in company with the labels of the training set. The labels of the test set are hidden at the time of writing this paper as they will be used for benchmarking machine learning algorithms on an open platform.

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[1–7]. Leveraging deep convolutional neural networks (CNN)[8], 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) physician's time is precious; (3) the annotation of medical images requires a consensus of several expert readers to overcome human bias[9]; 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-ray14[10], Padchest[11], CheXpert[3], and MIMIC-CXR[12]. 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 labels[13]. 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 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 JSRT[14], Indiana[15], MC[16], and SH[16].

Table 1. An overview of existing public datasets for CXR interpretation.

Dataset	Release year	# findings	# samples	Image-level labels	Local labels
JSRT[14]	2000	1	247$^{(\triangleleft,\star)}$	Available	Available
MC[16]	2014	1	138$^{(\triangleleft,\star)}$	Available	N/A
SH[16]	2014	1	662$^{(\triangleleft,\star)}$	Available	N/A
Indiana[15]	2016	10	8,121$^{(\triangleleft,\star)}$	Available	N/A
ChestX-ray8[10]	2017	8	108,948$^{(\bullet)}$	Available	Available$^{(\dagger)}$
ChestX-ray14[10]	2017	14	112,120$^{(\bullet)}$	Available	N/A
CheXpert[3]	2019	14	224,316$^{(\bullet)}$	Available	N/A
Padchest[11]	2019	193	160,868$^{(\bullet,\star)}$	Available	N/A$^{(\dagger\dagger)}$
MIMIC-CXR[12]	2019	14	377,110$^{(\bullet)}$	Available	N/A
VinDr-CXR (ours)	2020	28	18,000$^{(\star)}$	Available	Available

$^{(\bullet)}$ Labeled by an NLP algorithm. $^{(\star)}$ Labeled by radiologists. $^{(\triangleleft)}$ Moderate-size datasets that are not applicable for training deep learning models. $^{(\dagger)}$ A portion of the dataset (983 images) is provided with hand-labeled bounding boxes. $^{(\dagger\dagger)}$ 27% of the dataset was manually annotated with encoded anatomical regions of the findings.

Most of existing CXR datasets depend on automated rule-based labelers that either use keyword matching (e.g. CheXpert[3] and NIH labelers[10]) or an NLP model (e.g. CheXbert[17]) 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[13,18]. These noisy labels may lead to the deviation of deep learning-based algorithms from reported performances when evaluated in a real-world setting[19]. 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 (HMUH), 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 VinLab[1], which was built on top of a Picture Archiving and Communication System (PACS). All DICOM images and the labels of the training set are released. We temporarily retain the labels of the test set for the purpose of holding a CXR analysis competition on an open platform, which is expected to launch in December of 2020.

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 Figure 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, HMUH 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, and Samsung. 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.

[1] A demonstration of this framework can be found here: https://vindr.ai/vinlab

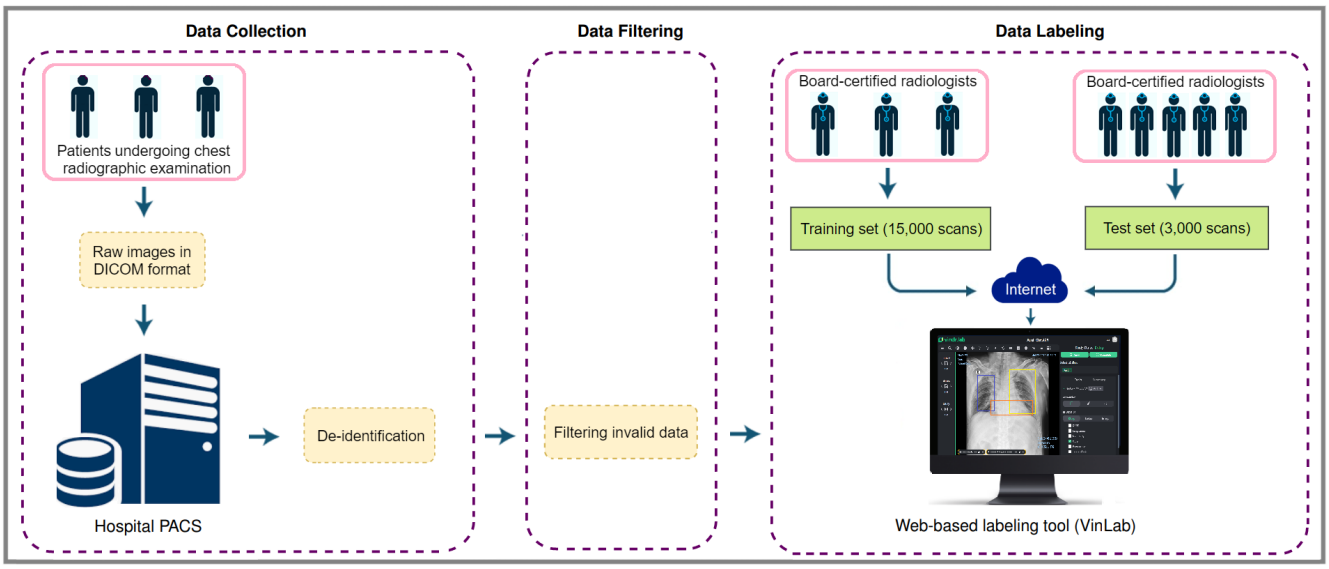

Figure 1. The flow of creating VinDr-CXR dataset: (1) raw images in DICOM format were collected retrospectively from the hospital's PACS and got de-identified to protect patient's privacy; (2) invalid files, such as images of other modalities, other body parts, low quality, or incorrect orientation, were automatically filtered out by a CNN-based classifier; (3) A web-based labeling tool, VinLab, 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.

Data de-identification

To protect patient's privacy[20], all personally identifiable information associated with the images has been removed or replaced with random values. Specifically, we ran a Python script that removes all DICOM tags of protected health information (PHI)[21] 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 and for clinical reference, like patient's sex and age. The entire list of retained attributes is shown in Table 4 (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 Figure 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 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 3 (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 VinLab and had a team of 17 experienced radiologists remotely annotate the data. All the radiologists participating in the labeling process were certified in

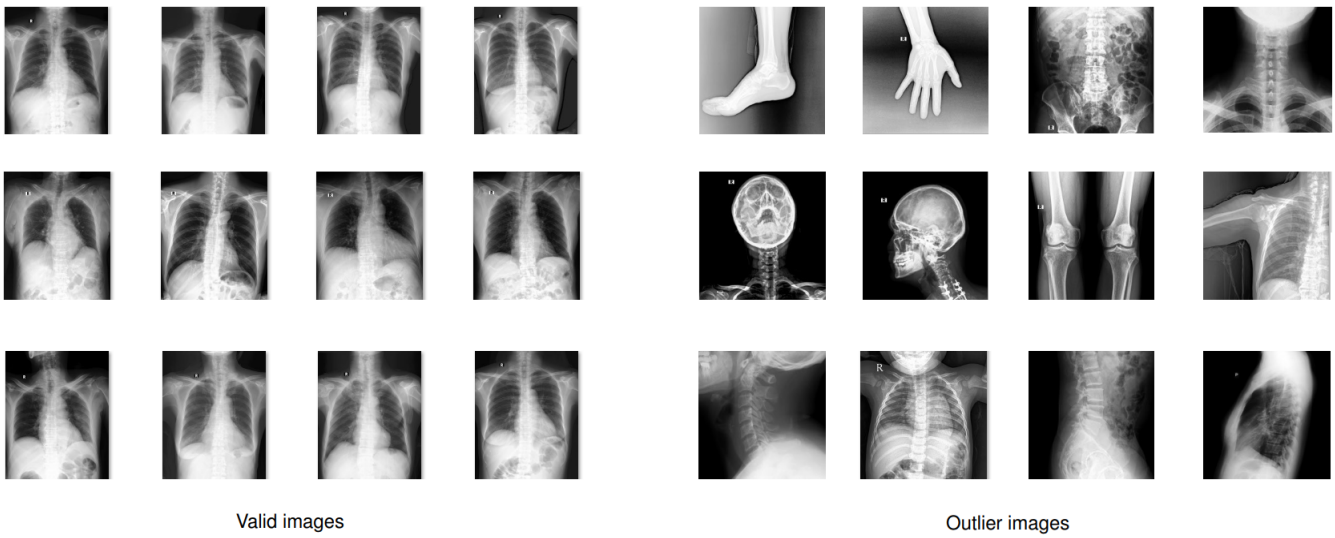

Figure 2. Examples of valid (**left**) and invalid (**right**) CXR scans. A CNN-based classifier was trained and used to automatically filter outliers; only valid PA-view CXRs of adults were retained for labeling.

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 serve as the training set and the rest 3,000 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 involved into a two-stage labeling process. 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 were carefully discussed and resolved by the 2 reviewers. Finally, the consensus of their opinions will serve as reference ground-truth.

Once the labeling has been 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 with 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 Figure 4. We have released all images together with the labels of the training set, while the annotations of test images are retained for the purpose of public model benchmarking.

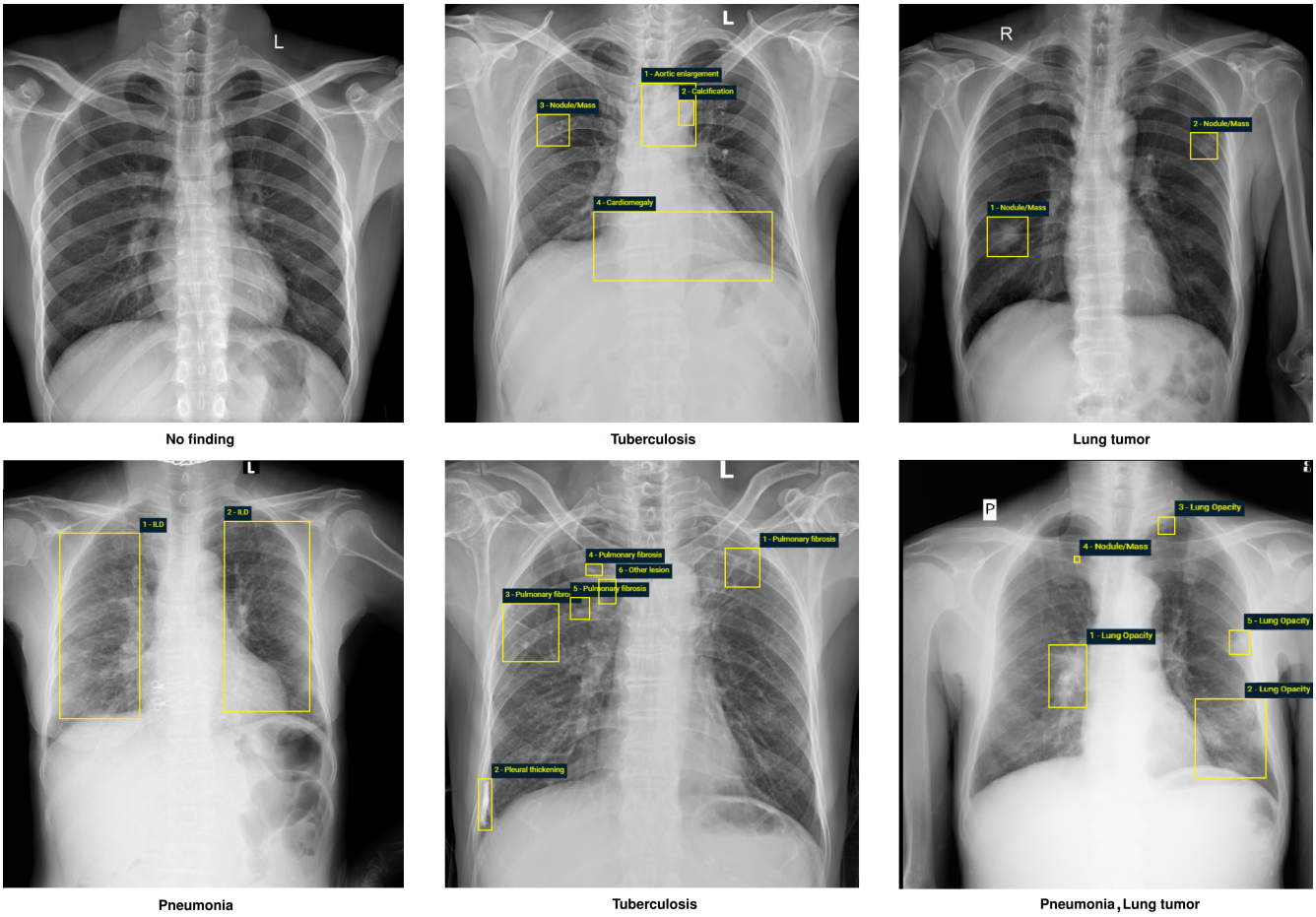

Figure 3. Examples of CXRs with radiologist's annotations. Abnormal findings (local labels) marked by radiologists are plotted on the original images for visualization purpose. The global labels are in bold and listed at the bottom of each example. Better viewed on a computer and zoomed in for details.

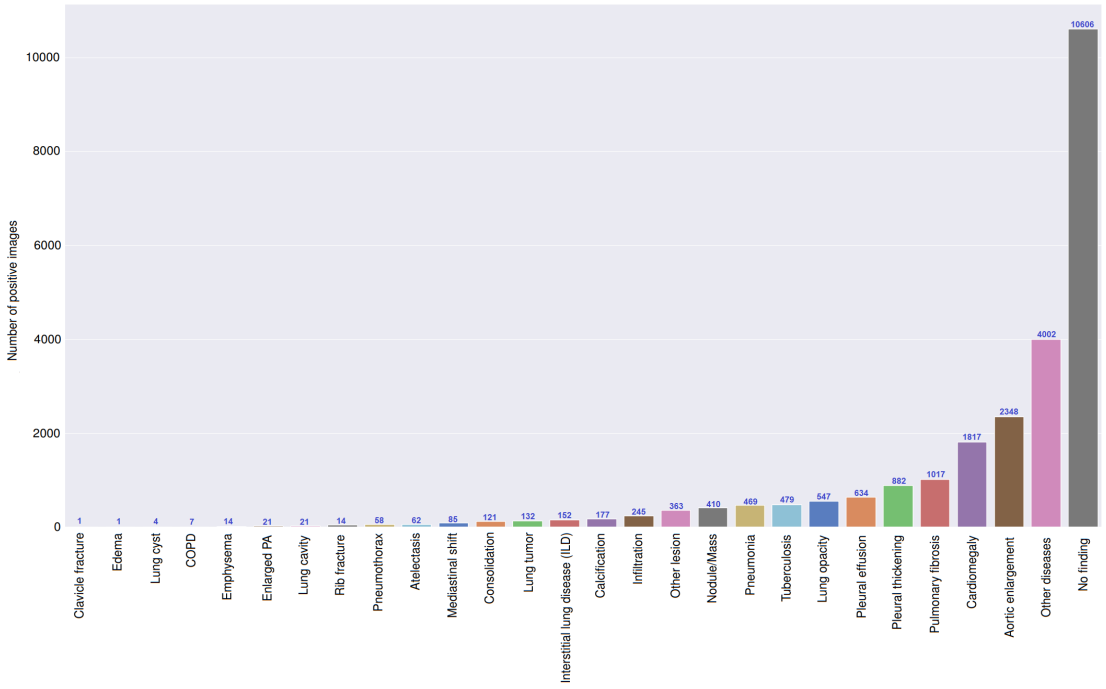

Figure 4. Distribution of findings and pathologies on the training set of VinDr-CXR.

Table 2. Dataset characteristics

	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%)	-
	2. Atelectasis (%)	62 (0.41%)	-
	3. Cardiomegaly (%)	1817 (12.11%)	-
	4. Calcification (%)	177 (1.18%)	-
	5. Clavicle fracture (%)	1 (0.01%)	-
	6. Consolidation (%)	121 (0.81%)	-
	7. Edema (%)	1 (0.01%)	-
	8. Emphysema (%)	14 (0.09%)	-
	9. Enlarged PA (%)	21 (0.14%)	-
	10. Interstitial lung disease (ILD) (%)	152 (1.01%)	-
	11. Infiltration (%)	245 (1.63%)	-
	12. Lung cavity (%)	21 (0.14%)	-
	13. Lung cyst (%)	4 (0.03%)	-
	14. Lung opacity (%)	547 (3.65%)	-
	15. Mediastinal shift (%)	85 (0.57%)	-
	16. Nodule/Mass (%)	410 (2.73%)	-
	17. Pulmonary fibrosis (%)	1017 (6.78%)	-
	18. Pneumothorax (%)	58 (0.39%)	-
	19. Pleural thickening (%)	882 (5.88%)	-
	20. Pleural effusion (%)	634 (4.23%)	-
	21. Rib fracture (%)	41 (0.27%)	-
	22. Other lesion (%)	363 (2.42%)	-
Global labels	23. Lung tumor (%)	132 (0.88%)	-
	24. Pneumonia (%)	469 (3.13%)	-
	25. Tuberculosis (%)	479 (3.19%)	-
	26. Other diseases (%)	4002 (26.68%)	-
	27. COPD (%)	7 (0.05%)	-
	28. No finding (%)	10606 (70.71%)	-

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. (-) To preserve the integrity of the test set, its labels are not released to the public at the time of writing this paper. The statistic of the labels on the test set is therefore not shown here.

Data Records

The VinDr-CXR dataset has been submitted to The Cancer Imaging Archive (TCIA) for public download. It is also accessible via our project website at https://vindr.ai/datasets/cxr/. We provide all imaging data and the corresponding ground truth labels for the training set only. 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 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 were stored in different CSV file, `image_labels_train.csv`, with the following fields: Image ID (`image_id`), radiologist ID (`rad_ID`), and global labels (`labels`). 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".

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[22], the European GDPR[23], as well as the local privacy laws[20]. 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 out-of-distribution 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 VinLab 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 register and accept a data use agreement (DUA) described on our webpage. 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. 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.

References

1. Rajpurkar, P. *et al.* CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).
2. 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).
3. 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*, vol. 33, 590–597 (2019).
4. 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).
5. Rajpurkar, P. *et al.* CheXpedition: Investigating generalization challenges for translation of chest X-ray algorithms to the clinical setting. *arXiv preprint arXiv:2002.11379* (2020).
6. Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Medicine* **3**, 1–8, https://doi.org/10.1038/s41746-020-0273-z (2020).
7. 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. *arXiv preprint arXiv:1911.06475* (2020).
8. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **512**, 436–444, https://doi.org/10.1038/nature14539 (2015).
9. 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).
10. 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).
11. 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. *arXiv preprint arXiv:1901.07441* (2019).
12. Johnson, A. E. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317, https://doi.org/10.1038/s41597-019-0322-0 (2019).
13. 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).

14. 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. *Am. J. Roentgenol.* **174**, 71–74, https://doi.org/10.2214/ajr.174.1.1740071 (2000).

15. Demner-Fushman, D. *et al.* Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Informatics Assoc.* **23**, 304–310, https://doi.org/10.1093/jamia/ocv080 (2016).

16. Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Medicine Surg.* **4**, 475–477, https://dx.doi.org/10.3978%2Fj.issn.2223-4292.2014.11.20 (2014).

17. Smit, A. *et al.* CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167* (2020).

18. Oakden-Rayner, L. Exploring large-scale public medical image datasets. *Acad. Radiol.* **27**, 106 – 112, https://doi.org/10.1016/j.acra.2019.10.006 (2020). Special Issue: Artificial Intelligence.

19. 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).

20. Vietnamese National Assembly. Regulation 40/2009/QH12 (Law on Medical Examination and Treatment). http://vbpl.vn/hanoi/Pages/vbpqen-toanvan.aspx?ItemID=10482 (2009). (Online; accessed 11 December 2020).

21. Isola, S. & Al Khalili, Y. Protected Health Information (PHI). https://www.ncbi.nlm.nih.gov/books/NBK553131/ (2019).

22. US Department of Health and Human Services. Summary of the HIPAA privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (2003).

23. European Parliament and Council of European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). https://gdpr-info.eu/ (2016). (Online; accessed 11 December 2020).

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.

Supplementary materials

Table 3. Definition of findings and diseases used in the study.

	Pathology label	Definition
Local Label	1. Aortic enlargement	An abnormal bulge that occurs in the wall of the major blood vessel.
	2. Atelectasis	Collapse of a part of the lung due to a decrease in the amount of air in the alveoli resulting in volume loss and increased density.
	3. Cardiomegaly	Enlargement of the heart, occurs when the heart of an adult patient is larger than normal and the cardiothoracic ratio is greater than 0.5.
	4. Calcification	Deposition of calcium salts in the lung.
	5. Clavicle fracture	A break in the collarbone.
	6. Consolidation	Any pathologic process that fills the alveoli with fluid, pus, blood, cells (including tumor cells) or other substances resulting in lobar, diffuse or multifocal ill-defined opacities.
	7. Edema	Fluid accumulation in the tissur and air space of the lungs.
	8. Emphysema	A condition of the lung characterized by an abnormal increase in the size of air spaces distal to the terminal bronchioles.
	9. Enlarged PA	Dilatation of the pulmonary artery - a defect characterized by a wider than normal main pulmonary artery.
	10. Interstitial lung disease (ILD)	Involvement of the supporting tissue of the lung parenchyma resulting in fine or coarse reticular opacities or small nodules.
	11. Infiltration	An abnormal substance that accumulates gradually within cells or body tissues or any substance or type of cell that occurs within or spreads as through the interstices (interstitium and/or alveoli) of the lung, that is foreign to the lung, or that accumulates in greater than normal quantity within it.
	12. Lung cavity	Thick-walled abnormal gas-filled spaces within the lung. They are usually associated with a nodule, mass, or area of consolidation. A fluid level within the space may be present.
	13. Lung cyst	Lung cysts refer to round, thin-walled, low attenuation spaces/lucencies in the lung.
	14. Lung opacity	Any abnormal focal or generalized opacity or opacities in lung fields (blanket tag including but not limited to consolidation, cavity, fibrosis, nodule, mass, calcification, interstitial thickening, etc.).
	15. Mediastinal shift	The deviation of the mediastinal structures towards one side of the chest cavity.
	16. Nodule/Mass	Any space occupying lesion either solitary or multiple.
	17. Pulmonary fibrosis	An excess of fibrotic tissue in the lung.
	18. Pneumothorax	The presence of gas (air) in the pleural space.
	19. Pleural thickening	Any form of thickening involving either the parietal or visceral pleura.
	20. Pleural effusion	Abnormal accumulations of fluid within the pleural space.
	21. Rib fracture	A common injury that occurs when one of the bones in your rib cage breaks or cracks.
	22. Other lesion	Other lesions that are not on the list of findings or abnormalities mentioned above.
Global labels	23. Lung tumor	The result of abnormal rates of cell division or cell death in lung tissue, or in the airways that lead to the lungs.
	24. Pneumonia	An infection that inflames the air sacs in one or both lungs.
	25. Tuberculosis	Any sign suggesting pulmonary or extrapulmonary tuberculosis.
	26. Other diseases	Other diseases that are not on the list of diseases mentioned above.
	27. COPD	Chronic obstructive pulmonary disease (COPD) is defined as a condition characterized by persistent airflow limitation that is usually progressive and associated with an enhanced chronic inflammatory response in the airways and the lung to noxious particles or gases.
	28. No finding	The absence of all pathologies from the chest radiograph.

Table 4. The list of DICOM tags that were retained for loading and processing raw images. All other tags were removed for protecting patient privacy. Details about all these tags can be found from DICOM Standard Browser at https://dicom.innolitics.com/ciods.

DICOM Tag	Attribute Name	Description
(0010, 0040)	Patient's Sex	Sex of the named patient.
(0010, 1010)	Patient's Age	Age of the patient.
(0010, 1020)	Patient's Size	Length or size of the patient, in meters.
(0010, 1030)	Patient's Weight	Weight of the patient, in kilograms.
(0028, 0010)	Rows	Number of rows in the image.
(0028, 0011)	Columns	Number of columns in the image.
(0028, 0030)	Pixel Spacing	Physical distance in the patient between the center of each pixel, specified by a numeric pair - adjacent row spacing (delimiter) adjacent column spacing in mm.
(0028, 0034)	Pixel Aspect Ratio	Ratio of the vertical size and horizontal size of the pixels in the image specified by a pair of integer values where the first value is the vertical pixel size, and the second value is the horizontal pixel size.
(0028, 0100)	Bits Allocated	Number of bits allocated for each pixel sample. Each sample shall have the same number of bits allocated.
(0028, 0101)	Bits Stored	Number of bits stored for each pixel sample. Each sample shall have the same number of bits stored.
(0028, 0102)	High Bit	Most significant bit for pixel sample data. Each sample shall have the same high bit.
(0028, 0103)	Pixel Representation	Data representation of the pixel samples. Each sample shall have the same pixel representation.
(0028, 0106)	Smallest Image Pixel Value	The minimum actual pixel value encountered in this image.
(0028, 0107)	Largest Image Pixel Value	The maximum actual pixel value encountered in this image.
(0028, 1050)	Window Center	Window center for display.
(0028, 1051)	Window Width	Window width for display.
(0028, 1052)	Rescale Intercept	The value b in relationship between stored values (SV) and the output units specified in Rescale Type (0028,1054). Each output unit is equal to m*SV + b.
(0028, 1053)	Rescale Slope	Value of m in the equation specified by Rescale Intercept (0028,1052).
(7FE0, 0010)	Pixel Data	A data stream of the pixel samples that comprise the image.
(0028, 0004)	Photometric Interpretation	Specifies the intended interpretation of the pixel data.
(0028, 2110)	Lossy Image Compression	Specifies whether an image has undergone lossy compression (at a point in its lifetime).
(0028, 2114)	Lossy Image Compression Method	A label for the lossy compression method(s) that have been applied to this image.
(0028, 2112)	Image Compression Ratio	Describes the approximate lossy compression ratio(s) that have been applied to this image.
(0028, 0002)	Samples per Pixel	Number of samples (planes) in this image.
(0028, 0008)	Number of Frames	Number of frames in a multi-frame image.