An Accurate and Explainable Deep Learning System Improves Interobserver Agreement in the Interpretation of Chest Radiograph

Hieu H. Pham[1,2,*,†], Ha Q. Nguyen[1,2,†], Khanh Lam[1,3,†], Linh T. Le[1,4,†], Dung B. Nguyen[1], Hieu T. Nguyen[1], Tung T. Le[1], Thang V. Nguyen[1], Minh Dao[1], and Van Vu[1,5]

[1]Medical Imaging Center, VinBigdata, Hanoi, Vietnam
[2]College of Engineering & Computer Science, VinUniversity, Hanoi, Vietnam
[3]Hospital 108, Department of Radiology, Hanoi, Vietnam
[4]Hanoi Medical University Hospital, Center of Radiology, Hanoi, Vietnam
[5]Yale University, Department of Mathematics, New Heaven, CT 06520, United States
[†]These authors contributed equally to this work
[*]Correspondence and requests for materials should be addressed to Hieu H. Pham (hieuhuy01@gmail.com)

ABSTRACT

Interpretation of chest radiographs (CXR) is a difficult but essential task for detecting thoracic abnormalities. Recent artificial intelligence (AI) algorithms have achieved radiologist-level performance on various medical classification tasks. However, only a few studies addressed the localization of abnormal findings from CXR scans, which is essential in explaining the image-level classification to radiologists. Additionally, the actual impact of AI algorithms on the diagnostic performance of radiologists in clinical practice remains relatively unclear. To bridge these gaps, we developed an explainable deep learning system called VinDr-CXR that can classify a CXR scan into multiple thoracic diseases and, at the same time, localize most types of critical findings on the image. VinDr-CXR was trained on 51,485 CXR scans with radiologist-provided bounding box annotations. It demonstrated a comparable performance to experienced radiologists in classifying 6 common thoracic diseases on a retrospective validation set of 3,000 CXR scans, with a mean area under the receiver operating characteristic curve (AUROC) of 0.967 (95% confidence interval [CI]: 0.958–0.975). The sensitivity, specificity, *F1*-score, false-positive rate (FPR), and false-negative rate (FNR) of the system at the optimal cutoff value were 0.933 (0.898–0.964), 0.900 (0.887–0.911), 0.631 (0.589–0.672), 0.101 (0.089–0.114) and 0.067 (0.057–0.102), respectively. For the localization task with 14 types of lesions, our free-response receiver operating characteristic (FROC) analysis showed that the VinDr-CXR achieved a sensitivity of 80.2% at the rate of 1.0 false-positive lesion identified per scan. A prospective study was also conducted to measure the clinical impact of the VinDr-CXR in assisting six experienced radiologists. The results indicated that the proposed system, when used as a diagnosis supporting tool, significantly improved the agreement between radiologists themselves with an increase of 1.5% in mean Fleiss' Kappa. We also observed that, after the radiologists consulted VinDr-CXR's suggestions, the agreement between each of them and the system was remarkably increased by 3.3% in mean Cohen's Kappa. Altogether, our results highlight the potentials of the proposed deep learning system as an effective assistant to radiologists in clinical practice. Part of the dataset used for developing the VinDr-CXR system has been made publicly available at `https://physionet.org/content/vindr-cxr/1.0.0/`.

Introduction

Common chest pathologies affect several hundred million people worldwide and kill several million cases every year[1,2]. They are the leading cause of death and impose an immense worldwide health burden[3]. Diagnosis of thoracic diseases is a crucial clinical task for physicians. Currently, chest X-ray (CXR) radiography is the primary imaging modality used for screening, triaging, and diagnosing varieties of lung conditions[4] such as pneumothorax, pneumonia, tuberculosis (TB), pleural effusion, atelectasis, emphysema, and cancers, etc. However, the CXR interpretation is a complicated task, which requires an in-depth understanding of radiologic signs in thoracic imaging[5–8]. A previous study[9] reported that 22% of all errors in diagnostic radiology were made in the CXR interpretation. A recent work[10] showed that 19%–26% of lung cancers visible on CXR images were missed at the first reading. Furthermore, interpreting CXR scans usually is highly dependent on the observer and has a poor interagreement between physicians[11]. The interobserver agreement was considered poor to moderate depending on the type of findings[12]; this rate can be lower in local hospitals, leading to unfavorable consequences.

Advanced machine learning algorithms have recently shown their significant potential in medical image analysis[13]. In particular, previous studies[14–23] have indicated that a deep learning (DL) system trained on a large-scale, annotated medical imaging dataset can reach a level of performance comparable to practicing radiologists in detecting common thoracic diseases[14–19], analyzing retinal images[20,21], or diagnosing skin cancers[22,23]. However, the actual impact of DL systems in clinical practice remains unclear, and large-scale clinical evaluations of these such systems are limited. Hence, despite many promising results and increasing performances that have been published, very few DL algorithms have reached clinical implementation. We observe that multiple factors slow or impede artificial intelligence (AI) transfer into clinical practice. First, the development of an accurate and robust DL system requires a large number of annotated CXR scans from diverse sources. The creation of large-scale, high-quality datasets of annotated images is costly and challenging. Meanwhile, public datasets are limited, and their labels are unreliable since they are produced by automated rule-based labelers[16,24,25]. Previous evidences[17,26,27] showed that training DL systems on small datasets and low-quality annotations raises concerns about the robustness of those systems in real clinical contexts. Second, few clinical evaluations of DL-assisted diagnostic algorithms have been performed, and most of them are not even prospective and at high risk of bias[28]. Third, although DL systems can outperform physicians in specific clinical tasks, the lack of explanatory power[29,30] became a key obstacle to convince medical experts, who must understand how and why DL models have made a prediction. We believe that an accurate assessment of CXRs requires both detection of abnormal findings and a correct decision at the disease level. Hence, the provision of accurate and interpretable visualizations of lung abnormalities is a crucial step towards the clinical translation of DL systems.

To address these gaps, we introduce in this study a fully automated DL system, namely VinDr-CXR, for chest radiograph interpretation. The VinDr-CXR is designed to simultaneously classify six common lung diseases and localize 14 important findings from CXR scans. The development and evaluation of VinDr-CXR are based on large-scale medical imaging analysis and state-of-the-art DL algorithms. Specifically, we use a patient dataset from multiple hospitals in Vietnam, containing 51,485 manually annotated CXR studies, to train the DL system. To evaluate the performance of the proposed system, we compare the model's performance with that of human experts in a benchmark study using the consensus annotations provided by 5 experienced radiologists on a retrospective dataset as the reference standard. To validate the generalization capability of VinDr-CXR, we compute its performance on various external datasets. The results confirm that our framework is accurate and robust across multiple populations and settings. To demonstrate the clinical value of VinDr-CXR as an assistant to radiologists, we conduct a prospective study at two hospitals in Vietnam and investigate the inter-rater agreement on CXR interpretation with and without VinDr-CXR's assistance. We further calculate the change in the agreement rate between VinDr-CXR itself and each radiologist before and after he/she consults the system's suggestions. Experiments indicate that the proposed DL system provides meaningful supports for radiologists in detecting thoracic diseases in a real-world clinical setting.

Results

This section provides an overview of the proposed approach and summarizes our main findings in this study. Please refer to the Methods and Supplementary Materials (pp 16–20) sections for complete details regarding the network architectures, model training procedure, and reader study design.

Overview of approach

We present in this section VinDr-CXR, a DL-based framework for the per-radiograph classification of common lung diseases and the detection of abnormal findings on CXRs. This framework includes two major components. First, an image-level classification network accepts a CXR scan as input and predicts whether it could be normal or abnormal. Second, a lesion-level detection network receives an abnormal CXR scan as input from the classifier and provides the location of abnormal findings via bounding box predictions. An overview of the proposed approach is illustrated in Figure 1A. The core of the VinDr-CXR system is based on state-of-the-art DL networks for image classification and object detection tasks, named EfficientNet[31] and EfficientDet[32], respectively. The classification network is trained with image-level labels to distinguish six common lung diseases, including pneumonia, tuberculosis, lung tumor, pleural effusion, other diseases, and no finding. Meanwhile, the detection network is trained with lesion-level annotations to localize 14 critical findings from the CXR images, i.e., cardiomegaly, opacity, consolidation, atelectasis, pneumothorax, pleural effusion, aortic enlargement, interstitial lung disease (ILD), infiltration, nodule/mass, pulmonary fibrosis, pleural thickening, calcification, and other lesions. To develop DL algorithms, a total of 51,485 anteroposterior (AP) and posteroanterior (PA) CXR scans have been retrospectively collected from the Hanoi Medical University Hospital (HMUH) and Hospital 108 (H108), which then manually annotated by expert radiologists. To evaluate the diagnostic accuracy of the VinDr-CXR, we compare its prediction with ground truth on an internal validation set of 3,000 studies, which was separate from the training set. Additionally, external datasets including CheXpert (N = 200) CheXphoto (N = 200) are used to verify the robustness of the VinDr-CXR for cross-site validation. Finally, we investigate the actual impact of the VinDr-CXR on clinical practice through a large reader study (N = 400). The inter-rater agreement among radiologists as well as the rate of agreement between VinDr-CXR and radiologists are then be assessed.

Figure 1B shows an overview of the development and evaluation of the VinDr-CXR framework. Full details are described in Supplementary Materials (pp 16–17).

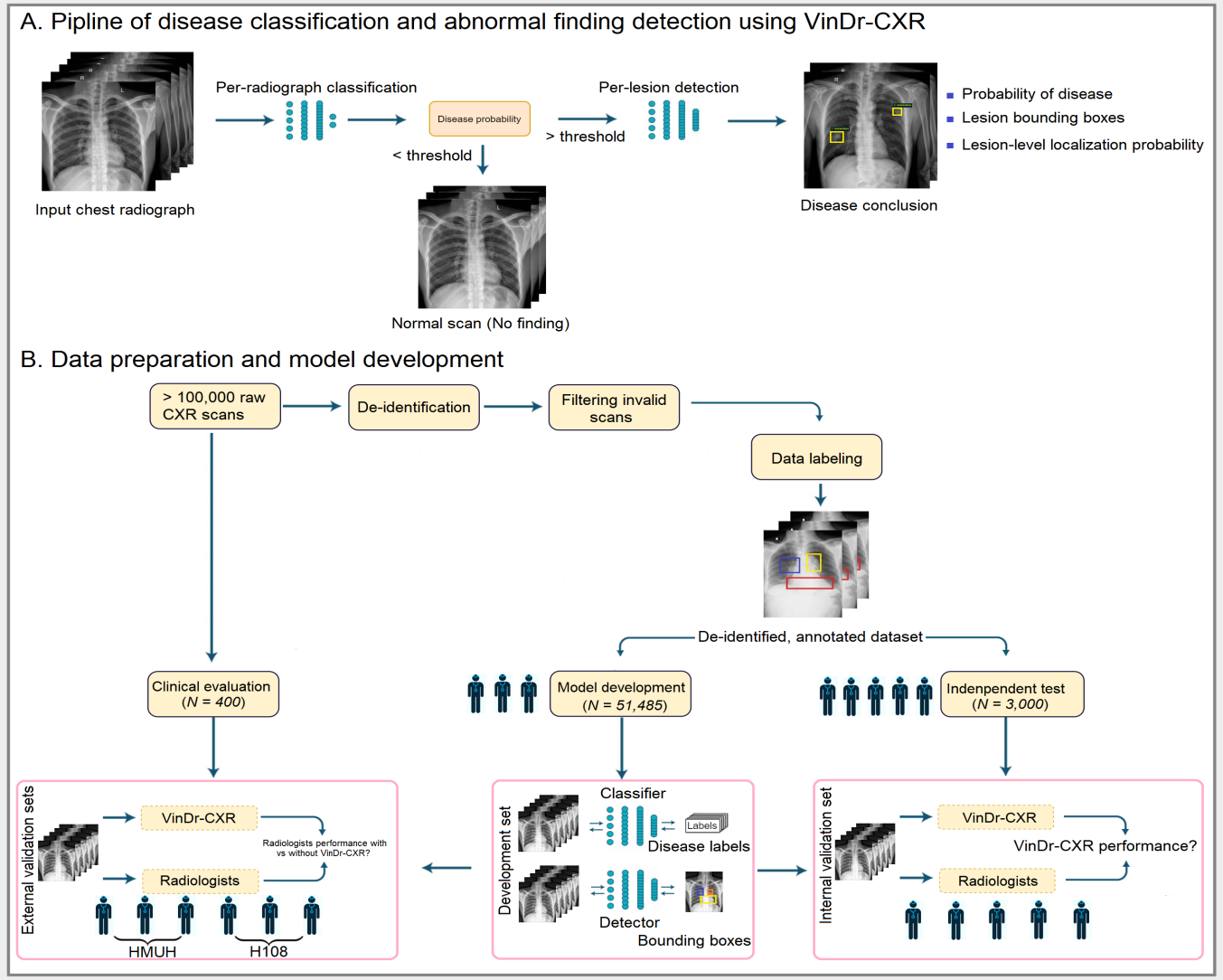

Figure 1. Overview of the proposed approach. **(A) The pipeline of the VinDr-CXR framework for lung disease classification and localization using CXR images**. A classifier takes as input one CXR scan and predicts its probability of abnormality. A detector takes an abnormal scan as input and provides bounding boxes with probabilities of abnormal findings at the lesion-level. The final prediction of the whole framework comes up with a complete description consisting of a disease conclusion, and abnormal findings. **(B) Summary of the VinDr-CXR development and evaluation**. The raw images in Digital Imaging and Communications in Medicine (DICOM) standard were collected retrospectively from the picture archiving and communication system (PACS) of the HMUH and H108 hospitals. The raw images were then de-identified and all out-of-distribution samples were removed. An in-house web-based labeling tool called VinDr Lab was designed to annotate imaging data. In the labeling process, each CXR scan from the development set was labeled by a group of three experienced radiologists for the presence of 22 preliminary findings and six disease impressions. Meanwhile, CXR scans from the internal validation set were annotated by a consensus of five experienced radiologists. Finally, 51,485 CXR images were used to develop DL algorithms, and 3,000 studies were used for validation. External evaluations were also performed on CheXpert[16] (*N = 200*) and CheXphoto[33] (*N = 200*) datasets to evaluate the robustness of the VinDr-CXR across multiple hospitals. A reader study was designed and performed at the HMUH and H108 (*N = 400*) to investigate the impact of the VinDr-CXR in clinical practice.

Table 1. Per-radiograph classification performance of the VinDr-CXR on the internal validation set (*N = 3,000*)

Dataset and Label	AUC	Sensitivity	Specificity	*F1*-score	FPR	FNR
HMUH & H108						
Pleural Effusion	.989 (.983, .994)	.955 (.912, .991)	.934 (.925, .943)	.524 (.463, .582)	.066 (.057, .075)	.045 (.009, .088)
Lung Tumor	.978 (.965, .988)	.937 (.877, .986)	.937 (.928, .946)	.448 (.381, .512)	.063 (.054, .072)	.063 (.014, .123)
Pneumonia	.969 (.959, .978)	.959 (.933, .982)	.877 (.864, .889)	.576 (.535, .616)	.123 (.111, .136)	.041 (.018, .067)
Tuberculosis	.975 (.964, .983)	.903 (.854, .945)	.936 (.927, .945)	.602 (.550, .651)	.064 (.055, .073)	.097 (.055, .146)
Other Diseases	.920 (.909, .931)	.925 (.904, .945)	.796 (.780, .812)	.698 (.674, .722)	.204 (.188, .220)	.075 (.055, .096)
No Finding	.972 (.966, .978)	.920 (.908, .932)	.913 (.895, .930)	.939 (.931, .947)	.087 .070, .105)	.080 (.068, .092)
Mean	**.967 (.958, .975)**	**.933 (.898, .964)**	**.900 (.887, .911.)**	**.631 (.589, 672.)**	**.101 (.089, .114)**	**.067 (.057, .102)**

Abbreviations: AUC = Area under the receiver operating characteristic curve, FPR = False-positive rate or false alarm rate, FNR = False-negative rate or miss detection rate.

Table 2. Per-radiograph classification performance of the VinDr-CXR on Pleural Effusion, Pneumonia and No Finding from the external validation sets CheXpert[16] (*N = 200*) and CheXphoto[33] (*N = 200*) datasets

Dataset and Label	AUC	Sensitivity	Specificity	*F1*-score	FPR	FNR
CheXpert[16]						
Pleural Effusion	.895 (.850, .934)	.940 (.877, .987)	.727 (.657, .795)	.715 (.637, .788)	.273 (.205, .343)	.060 (.013, .123)
Pneumonia	.891 (.824, .949)	1.00 (1.00, 1.00)	.248 (.193, .306)	.085 (.033, .143)	.752 (.694, .807)	.000 (.000, .000)
No Finding	.891 (.843, .933)	.659 (.500, .810)	.868 (.818, .914)	.559 (.424, .678)	.132 (.086, .182)	.341 (.190, .500)
Mean	**.892 (.839, .939)**	**.866 (.792, .932)**	**.614 (.556, .672)**	**.453 (.365, .536)**	**.386 (.328, .444)**	**.134 (.067, .208)**
CheXphoto[33]						
Pleural Effusion	.889 (.843, .930)	.955 (.899, 1.00)	.739 (.670, .804)	.731 (.654, .802)	.261 (.196, .331)	.045 (.000, .101)
Pneumonia	.887 (.816, .947)	1.00 (1.00, 1.00)	.230 (.176, .288)	.083 (.032, .140)	.770 (.712, .824)	.000 (.000, .000)
No Finding	.887 (.836, .930)	.658 (.500, .808)	.873 (.824, .919)	.566 (.430, .684)	.127 (.081, .176)	.342 (.194, .500)
Mean	**.888 (.832, .936)**	**.871 (.800, .936)**	**.614 (.557, .670)**	**.460 (.372, .542)**	**.386 (.330, .444)**	**.129 (.065, .200)**

Abbreviations: AUC = Area under the receiver operating characteristic curve, FPR = False-positive rate or false alarm rate, FNR = False-negative rate or miss detection rate.

Evaluation of VinDr-CXR performance

The following subsections detail the quantitative results of the VinDr-CXR system for the classification of diseases and detection of critical lesions on internal and external validation datasets.

VinDr-CXR provides accurate per-radiograph classification of common lung disease

The performance of the VinDr-CXR for the classification of common lung diseases was assessed on the internal validation set of 3,000 CXR studies, in which there were 948 patients with abnormal findings or diseases and 2052 patients without any pathologies (Table 8). The system achieved a mean AUROC of 0.967 (95% CI: 0.958, 0.975) over six global disease labels: 0.989 (0.983, 0.994) for Pleural Effusion, 0.978 (0.965, 0.988) for Lung Tumor, 0.969 (0.959, 0.978) for Pneumonia, 0.975 (0.964, 0.983) for Tuberculosis, and 0.920 (0.909, 0.931) for Other Diseases. The sensitivity, specificity, and *F1*-score of the VinDr-CXR were 0.933 (0.898, 0.964), 0.900 (0.887, 0.911), and 0.631 (0.589, 0.672), respectively. The system showed a FPR of 0.101 (0.089, 0.114) and a FNR of 0.067 (0.057, 0.102) over all target diseases. The overall accuracies of the system in differentiating between normal and abnormal CXRs were 0.972 (0.966, 0.978) in AUROC and 0.939 (0.931, 0.947) in *F1*-score. Our experimental results on CXR data from the internal validation set showed high sensitivity and specificity in classifying six disease labels. Detailed performances for individual diseases over all evaluation metrics on the internal validation cohort are reported in Table 1. Figure 2 shows six ROC curves of the system on the internal validation set for six global diseases.

VinDr-CXR shows robust classification performance on external test cohorts

To investigate the consistency of the VinDr-CXR performance across multiple populations, we performed external validation tests using two independent datasets, including CheXpert[16] and CheXphoto[33]. The AUROC score, sensitivity, specificity, *F1*-score, FPR, and FNR of the VinDr-CXR with no additional training cost on 200 studies [normal: 27 cases, abnormal: 173 cases] of the CheXpert[16] validation set were 0.892 (0.839, 0.939), 0.866 (0.792, 0.932), 0.614 (0.556, 0.672), 0.453 (0.3650, 0.536), 0.386 (0.328, 0.444), and 0.134 (0.067, 0.208), respectively. On the CheXphoto[33] validation set [*N = 200*, normal: 27 cases, abnormal: 173 cases], the VinDr-CXR achieved an AUC of 0.888 (0.832, 0.936), a sensitivity of 0.871 (0.800, 0.936), a specificity of 0.614 (0.557, 0.670), a *F1*-score of 0.460 (0.372, 0.542), a FPR of 0.386 (0.330, 0.444), and a FNR of 0.129 (0.065, 0.200). The performance of the VinDr-CXR over all disease labels is reported in Table 2. Although a slight drop has been observed, the performance of the VinDr-CXR on external test sets remained at a high level, showing its robustness across different patient cohorts. These experimental results show evidence that training a DL system with a large-scale, high-quality dataset could reach a high diagnostic accuracy across populations without additional training cost.

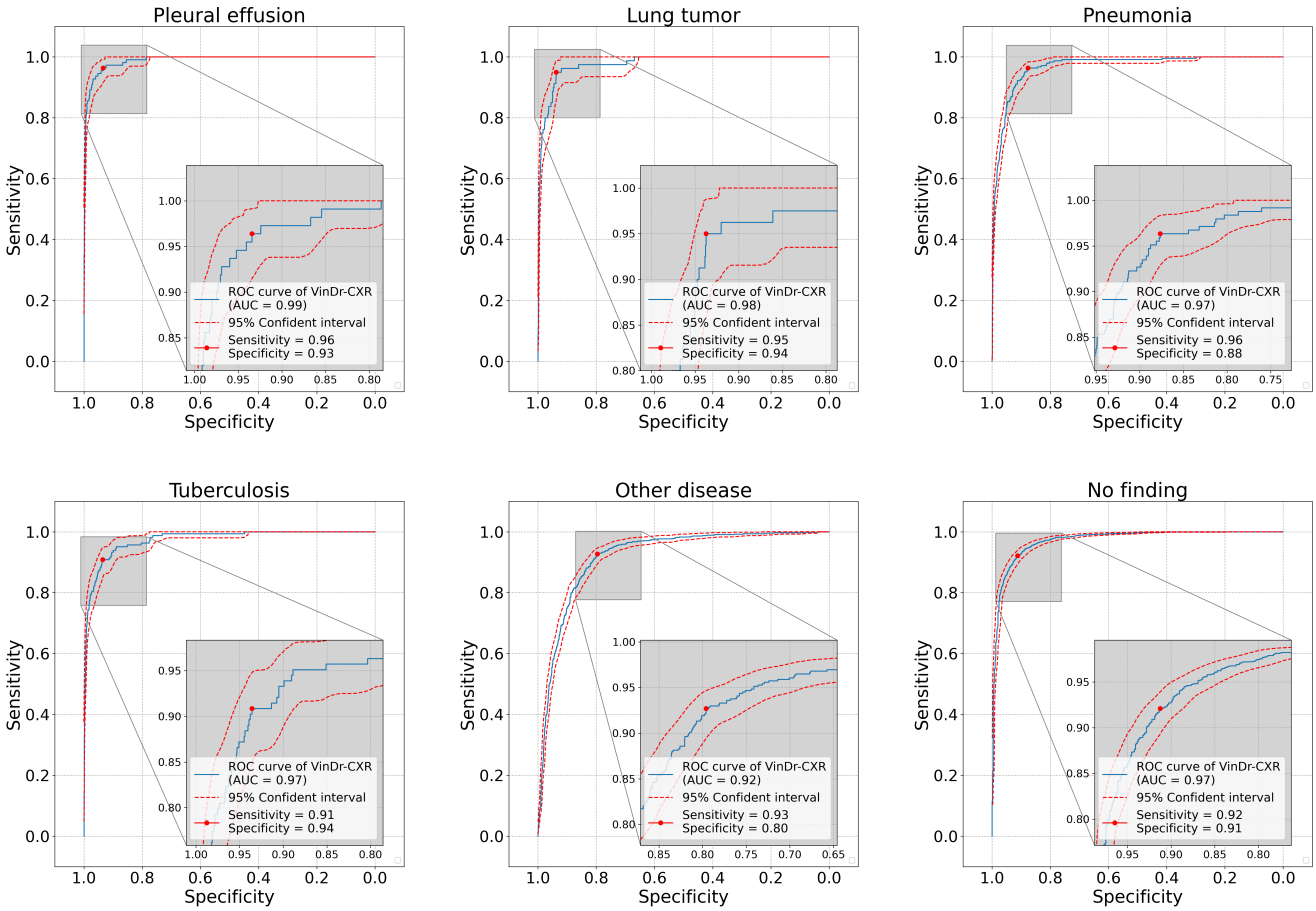

Figure 2. Receiver operating curves (ROC) of the VinDr-CXR system on the internal validation cohort. The solid blue lines show the ROC curves of the system and the dashed red lines show the 95 percentile intervals of the curves based on 10,000 bootstrap samples. We determined the optimal threshold for the VinDr-CXR system by maximizing Youden's index[34] for each disease label. We observe that the DL system achieved consistently high classification performances across all the target diseases (range AUROC: 0.920 – 0.989).

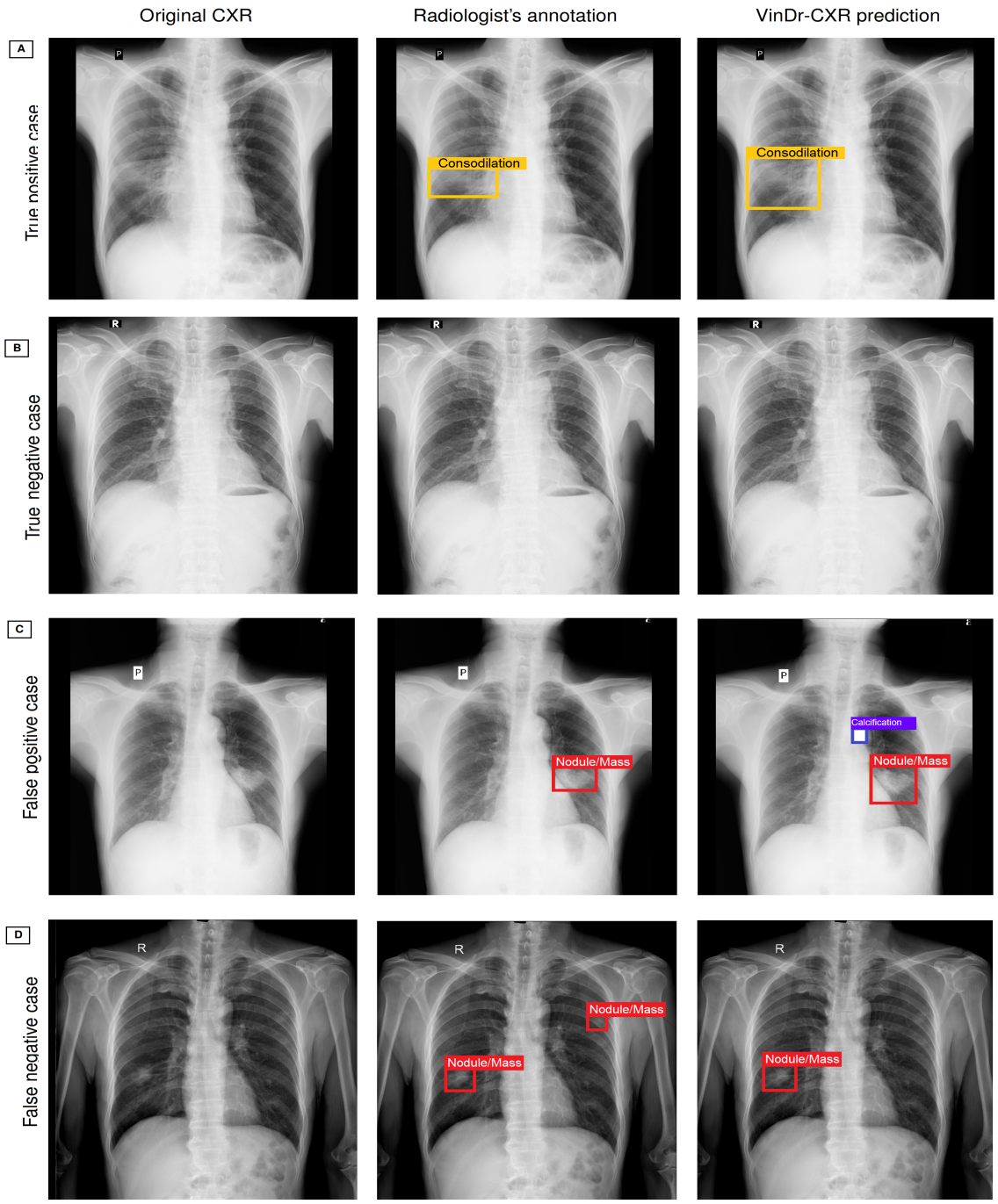

Figure 3. Some accurate and erroneous predictions of the VinDr-CXR on several representative CXR images from the internal validation set. (**A**) The VinDr-CXR correctly identified four lesions, including Pleural Thickening, Calcification, Aortic Enlargement, and Cardiomegaly on the scan. (**B**) The system correctly identified a normal scan or true negative case. (**C**) The VinDr-CXR correctly identified a Nodule/Mass. However, it missed another Nodule/Mass on the left lung. (**D**) The VinDr-CXR correctly identified Nodule/Mass, but it seemed wrong to detect Calcification.

Table 3. Abnormality detection performance of the VinDr-CXR on the internal validation dataset

Finding	Sensitivity (δ = 0.25)	Sensitivity (δ = 0.50)	Sensitivity (δ = 1.00)	Sensitivity (δ = 2.00)	Sensitivity (δ = 4.00)
Cardiomegaly	.965 (.943, .984)	.965 (.943, .984)	.968 (.947, .986)	.968 (.947, .986)	.968 (.947, .986)
Opacity	.617 (.517, .718)	.756 (.660, .848)	.842 (.764, .915)	.895 (.827, .956)	.906 (.848, .959)
Consolidation	.841 (.764, .912)	.898 (.834, .954)	.937 (.884, .980)	.937 (.884, .980)	.937 (.884, .980)
Atelectasis	.642 (.538, .747)	.698 (.596, .796)	.772 (.678, .859)	.772 (.678, .859)	.772 (.678, .859)
Pneumothorax	.639 (.467, .800)	.680 (.522, .828)	.680 (.522, .828)	.680 (.522, .828)	.680 (.522, .828)
Pleural Effusion	.898 (.840, .950)	.927 (.877, .970)	.934 (.891, .972)	.942 (.902, .977)	.942 (.902, .977)
Aortic Enlargement	.838 (.789, .885)	.882 (.839, .923)	.905 (.864, .941)	.905 (.864, .942)	.909 (.870, .945)
Interstitial Lung Disease	.664 (.606, .723)	.782 (.730, .833)	.858 (.816, .899)	.923 (.890, .954)	.937 (.905, .964)
Infiltration	.801 (.714, .884)	.861 (.787, .930)	.911 (.843, .969)	.911 (.843, .969)	.911 (.843, .969)
Nodule/Mass	.579 (.510, .647)	.663 (.601, .727)	.735 (.670, .796)	.769 (.710, .827)	.773 (.714, .830)
Pulmonary Fibrosis	.568 (.514, .623)	.627 (.576, .678)	.707 (.657, .757)	.775 (.729, .819)	.796 (.751, .839)
Pleural Thickening	.494 (.425, .564)	.608 (.540, .675)	.714 (.651, .776)	.797 (.739, .852)	.850 (.797, .900)
Calcification	.598 (.527, .670)	.685 (.613, .755)	.774 (.713, .833)	.802 (.744, .858)	.802 (.744, .858)
Other Lesions	.265 (.197, .338)	.311 (.238, .388)	.372 (.299, .448)	.484 (.408, .562)	.551 (.474, .627)
No Finding	.921 (.909, .933)	.921 (.909, .933)	.921 (.909, .933)	.921 (.909, .933)	.921 (.909, .933)
Mean	**.689 (.668, .710)**	**.751 (.731, .770)**	**.802 (.784, .819)**	**.832 (.814, .849)**	**.844 (.826, .860)**

Model performance was evaluated using the FROC score. Here, the symbol δ denotes the number of false-positive predictions per image. Data in parentheses are 95% confidence intervals.

VinDr-CXR provides accurate lesion-level localization

For the per-lesion localization task, the VinDr-CXR's ability to detect and localize abnormal findings was evaluated on 3,000 CXR scans from the internal validation set using FROC analysis[35]. In this experiment, a detection is considered a true positive if the detected bounding box overlaps with the corresponding ground-truth bounding box more than 40% using the intersection over union (IoU) metric. Otherwise, it is considered to be a false positive. As shown in Table 3, the proposed system achieved a sensitivity of 80.2% (81.4, 84.9) at 1.0 false-positive marks per image. The FROC (average recall rate at the false positives as 0.25, 0.5, 1.00, 2.00, and 4.00) of the VinDr-CXR system was 78.36% (76.46, 80.16). FROC curves, which show the sensitivity of the system as a function of the number of false positives marks per image, of some representative findings are shown in Figure 4.

Anomaly detection visualization

The per-lesion detection performance of the VinDr-CXR system can be inspected visually through Figure 3. We investigated the characteristics and errors of the VinDr-CXR detector by visualizing several representative cases containing both correctly detected lesions and lesions that the DL system missed. In this experiment, we used the VinDr-CXR at a sensitivity of 0.802 and 1.0 false-positive marks per image to generate predictions. We found that the system was able to correctly identified almost all critical lesions. Meanwhile, most false-positive detections were small and non-dangerous lung lesions such as calcifications.

Assessment of VinDr-CXR performance in clinical practice

Inter-agreement among radiologists with and without VinDr-CXR assistance

We conducted a multi-reader study at the HMUH and H108 to investigate the impact of the proposed DL system on radiologists' performance in the regular clinical workflow. At each hospital, three experienced radiologists assessed a total of 200 CXR images without VinDr-CXR assistance at the first read. The images were read with the assistance of VinDr-CXR in the second read. The agreement between the VinDr-CXR and radiologists, as well as inter-observer agreement among radiologists were assessed using the percentage agreement rate and Kappa statistics. The details of the proposed reader study design are provided in the Methods section. In the first read, concordance among three radiologists on data from the H108 showed percentage agreement rates between 89.9%–90.7% and Cohen's Kappa values between 0.501 (0.349, 0.654)–0.540 (0.392, 0.688). Agreement among three H108's radiologists was moderate with a Fleiss' Kappa of 0.529 (0.453, 0.605). In the second read, with the support of VinDr-CXR system, agreement among three raters showed a slightly higher percentage agreement rate. Specifically, percentage agreement rates ranged between 90.2%–90.7% and Cohen's Kappa values between 0.519 (0.351, 0.678)–0.556 (0.412, 0.700) that indicated moderate inter-individual agreement. Agreement among three H108's radiologists was moderate with a Fleiss' Kappa of 0.545 (0.465, 0.625), corresponding to an 3.0% improvement in Fleiss' Kappa compared to the first read.

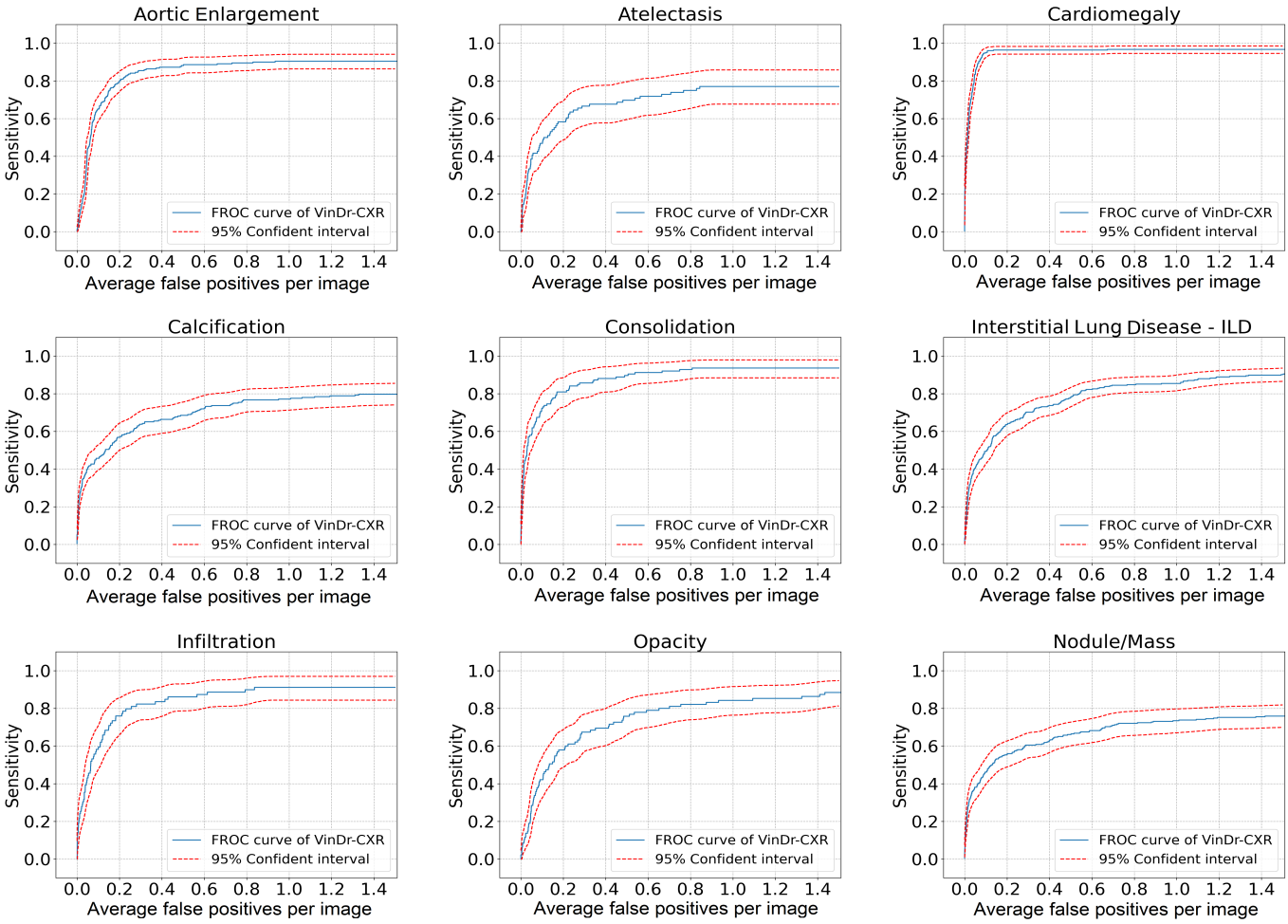

Figure 4. Free-response receiver operating characteristic (FROC) of the VinDr-CXR detector for several findings on the internal validation set. Sensitivity is calculated on a per lesion basis. The blue curve shows the results using all validation images, while the red dashed lines show the 95 percentile intervals from 10,000 bootstrap samples.

Table 4. Inter-rater agreement among radiologists at the H108 without the VinDr-CXR assistance

Findings	Rater 1 *vs.* Rater 2		Rater 1 *vs.* Rater 3		Rater 2 *vs.* Rater 3		Fleiss' κ
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ	
Lung Tumor	.920	.562 (.375, .748)	.955	.643 (.426, .859)	.905	.490 (.296, .683)	.552 (.472, .632)
Pneumonia	.895	.698 (.580, .817)	.875	.543 (.387, .700)	.880	.626 (.496, .756)	.626 (.546, .706)
Tuberculosis	.945	.674 (.494, .853)	.925	.596 (.413, .778)	.950	.763 (.623, .904)	.681 (.601, .761)
Other Diseases	.830	.651 (.544, .758)	.810	.595 (.483, .707)	.820	.617 (.507, .727)	.619 (.539, .699)
No Finding	.945	.886 (.821, .952)	.940	.878 (.812, .945)	.945	.889 (.825, .952)	.884 (.804, .964)
Aortic Enlargement	.915	.662 (.514, .811)	.895	.546 (.377, .716)	.930	.669 (.509, .830)	.625 (.545, .705)
Atelectasis	.925	-.009 (-.027, .008)	.980	-.008 (-.019, .004)	.935	.216 (-.046, .478)	.084 (.004, .164)
Calcification	.905	.654 (.512, .795)	.870	.255 (.058, .452)	.835	.288 (.122, .453)	.419 (.339, .499)
Cardiomegaly	.915	.574 (.393, .756)	.965	.810 (.673, .946)	.950	.734 (.578, .891)	.703 (.623, .783)
Consolidation	.940	.308 (.026, .591)	.940	.506 (.264, .749)	.940	.308 (.026, .591)	.388 (.308, .468)
ILD	.845	.552 (.412, .692)	.820	.348 (.189, .507)	.775	.196 (.044, .349)	.372 (.292, .452)
Infiltration	.815	.288 (.139, .436)	.860	.239 (.052, .426)	.865	.562 (.419, .706)	.376 (.296, .456)
Lung Opacity	.830	.167 (-.014, .347)	.870	.054 (-.067, .175)	.890	-.019 (-.042, .005)	.072 (-.008, .152)
Nodule/Mass	.895	.512 (.331, .694)	.910	.521 (.328, .715)	.915	.591 (.417, .766)	.541 (.461, .621)
Pleural Effusion	.930	.793 (.690, .896)	.905	.682 (.549, .815)	.935	.800 (.697, .903)	.760 (.680, .840)
Pleural Thickening	.835	.284 (.116, .451)	.890	.255 (.037, .473)	.865	.458 (.291, .624)	.337 (.257, .417)
Pneumothorax	1.00	1.00 (1.00, 1.00)	1.00	1.00 (1.00, 1.00)	1.00	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
Pulmonary Fibrosis	.870	.505 (.340, .669)	.890	.514 (.337, .691)	.900	.599 (.441, .756)	.539 (.458, .619)
Other Lesions	.920	.386 (.145, .627)	.935	.546 (.327, .765)	.925	.476 (.250, .702)	.471 (.391, .551)
Mean	**.899**	**.534 (.389, .679)**	**.907**	**.501 (.349, .654)**	**.903**	**.540 (.392, .688)**	**.529 (.453, .605)**

Table 5. Inter-rater agreement among the H108's radiologists with the VinDr-CXR assistance

Findings	Rater 1 *vs.* Rater 2		Rater 1 *vs.* Rater 3		Rater 2 *vs.* Rater 3		Fleiss' κ
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ	
Lung Tumor	.915	.525 (.333, .717)	.960	.671 (.458, .884)	.895	.436 (.239, .632)	.524 (.444, .604)
Pneumonia	.920	.778 (.675, .881)	.905	.690 (.561, .818)	.885	.655 (.529, .781)	.709 (.629, .789)
Tuberculosis	.950	.710 (.541, .879)	.930	.646 (.477, .814)	.950	.772 (.637, .907)	.711 (.631, .791)
Other Diseases	.815	.621 (.511, .731)	.810	.594 (.480, .707)	.825	.633 (.526, .740)	.614 (.534, .694)
No finding	.950	.897 (.834, .959)	.945	.889 (.825, .952)	.945	.889 (.825, .952)	.891 (.811, .971)
Aortic Enlargement	.920	.687 (.543, .830)	.895	.559 (.393, .726)	.935	.708 (.559, .857)	.650 (.570, .730)
Atelectasis	.935	.125 (-.098, .348)	.975	-.008 (-.021, .005)	.940	.312 (.033, .590)	.185 (.105, .265)
Calcification	.930	.764 (.646, .881)	.835	.192 (.018, .365)	.825	.268 (.105, .430)	.438 (.358, .518)
Cardiomegaly	.925	.639 (.471, .807)	.970	.833 (.703, .963)	.955	.775 (.635, .916)	.745 (.665, .825)
Consolidation	.930	.197 (-.056, .451)	.935	.484 (.243, .724)	.935	.211 (-.055, .477)	.320 (.240, .400)
ILD	.845	.552 (.412, .692)	.820	.373 (.214, .533)	.795	.296 (.137, .454)	.411 (.331, .491)
Infiltration	.810	.276 (.127, .425)	.835	.229 (.059, .399)	.885	.646 (.514, .777)	.402 (.322, .482)
Lung Opacity	.830	.167 (-.014, .347)	.885	.262 (.063, .462)	.865	-.055 (-.086, -.023)	.131 (.051, .211)
Nodule/Mass	.895	.557 (.389, .726)	.905	.573 (.401, .745)	.920	.647 (.487, .806)	.592 (.512, .672)
Pleural Effusion	.940	.825 (.730, .920)	.925	.759 (.643, .875)	.945	.834 (.739, .928)	.807 (.727, .887)
Pleural Thickening	.840	.318 (.151, .485)	.885	.197 (-.015, .410)	.855	.420 (.253, .587)	.321 (.241, .401)
Pneumothorax	.995	.798 (.410, 1.00)	.995	.798 (.410, 1.00)	1.00	1.00 (1.00, 1.00)	.855 (.775, .935)
Pulmonary Fibrosis	.865	.518 (.360, .675)	.875	.530 (.368, .691)	.900	.661 (.525, .798)	.571 (.491, .651)
Other Lesions	.920	.386 (.145, .627)	.940	.593 (.385, .801)	.920	.457 (.233, .681)	.482 (.402, .562)
Mean	**.902**	**.544 (.374, .704)**	**.907**	**.519 (.351, .678)**	**.904**	**.556 (.412, .700)**	**.545 (.465, .625)**

We observed the same results on the data from the HMUH. Details of clinical evaluation results at the HMUH are provided in Table 10 and Table 11 (Supplementary Materials, pp 17–18). In the first read, the percentage agreement rates between each pair of radiologists ranged from 90.2%–90.8%; Cohen's Kappa values were between 0.367 (0.192, 0.543)–0.483 (0.304, 0.662), and a Fleiss' Kappa of 0.404 (0.329–0.480). In the second read, percentage agreement rates were between 90.2%–90.8%, and Cohen's Kappa values were between 0.367 (0.192, 0.543)–0.483 (0.304, 0.662). The agreement between three HMUH's radiologists was a Fleiss' Kappa of 0.418 (0.342, 0.494), corresponding to an improvement of 3.4% compared to the first read.

Diagnostic agreement between VinDr-CXR and radiologists with and without assistance

In this experiment, the rate of agreement between the VinDr-CXR and radiologists in detecting abnormal lung findings from CXRs were assessed. Table 6 and Table 7 show the agreement rate between the VinDr-CXR and H108's radiologists without and with assistance, respectively. In the first read, the percentage agreement rates were 88.8%–90.2%, and the Cohen's Kappa values ranged from 0.462 (0.283, 0.640)–0.506 (0.327, 0.685). With the assistance of VinDr-CXR, the rates of agreement ranged from 90.5%–91.1% in percentage agreement and from 0.524 (0.348, 0.699)–0.546 (0.370, 0.717) in Cohen's Kappa values. At the HMUH, the percentage agreement rates were 88.8%–90.2%, and the Cohen's Kappa values ranged from 0.462 (0.283, 0.640)–0.506 (0.327, 0.685) without the VinDr-CXR assistance. Meanwhile, the rates of agreement ranged from 90.5%–91.1% in percentage agreement and from 0.524 (0.348, 0.699)–0.546 (0.370, 0.717) in Kappa values with the VinDr-CXR assistance (Detailed in Table 12 and Table 13 in the Supplementary Materials, p 18). After consulting the VinDr-CXR output, significant improvements in Kappa scores have been observed across two hospitals. These results indicated that the VinDr-CXR assistance resulted in a significant increase in the agreement between the DL system and radiologists.

Table 6. Agreement between the VinDr-CXR and H108's radiologists without assistance

Findings	Rater 1 *vs.* AI		Rater 2 *vs.* AI		Rater 3 *vs.* AI	
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ
Lung Tumor	.945	.493 (.235, .751)	.885	.329 (.129, .530)	.940	.469 (.215, .723)
Pneumonia	.900	.688 (.561, .815)	.905	.735 (.623, .846)	.875	.566 (.418, .715)
Tuberculosis	.965	.769 (.604, .934)	.950	.723 (.560, .885)	.940	.695 (.534, .856)
Other Diseases	.725	.399 (.277, .521)	.705	.355 (.231, .479)	.795	.487 (.353, .620)
No Finding	.905	.810 (.730, .890)	.900	.800 (.718, .882)	.935	.870 (.802, .938)
Aortic Enlargement	.880	.359 (.174, .544)	.895	.361 (.159, .563)	.915	.378 (.150, .605)
Atelectasis	.910	.091 (-.076, .259)	.895	.308 (.087, .529)	.900	.067 (-.093, .226)
Calcification	.875	.326 (.127, .525)	.860	.418 (.252, .585)	.875	.008 (-.136, .152)
Cardiomegaly	.920	.607 (.432, .782)	.905	.542 (.360, .724)	.935	.662 (.491, .832)
Consolidation	.920	.458 (.235, .681)	.910	.219 (-.006, .444)	.930	.526 (.309, .743)
ILD	.830	.488 (.340, .636)	.845	.537 (.394, .680)	.840	.385 (.219, .550)
Infiltration	.830	.226 (.062, .391)	.905	.712 (.592, .833)	.870	.530 (.372, .687)
Lung Opacity	.860	.338 (.147, .528)	.860	.255 (.058, .453)	.880	-.019 (-.043, .005)
Nodule/Mass	.860	.383 (.200, .566)	.875	.489 (.316, .661)	.870	.409 (.225, .593)
Pleural Effusion	.975	.921 (.853, .989)	.945	.839 (.747, .931)	.920	.735 (.613, .857)
Pleural Thickening	.905	.337 (.109, .565)	.850	.390 (.220, .559)	.915	.494 (.286, .701)
Pneumothorax	.985	.565 (.125, 1.00)	.985	.565 (.125, 1.00)	.985	.565 (.125, 1.00)
Pulmonary Fibrosis	.900	.558 (.386, .730)	.850	.398 (.224, .572)	.870	.384 (.194, .575)
Other Lesions	.945	.563 (.332, .794)	.955	.643 (.426, .859)	.940	.568 (.350, .787)
Mean	**.897**	**.494 (.308, .679)**	**.888**	**.506 (.327, .685)**	**.902**	**.462 (.283, .640)**

Table 7. Agreement between the VinDr-CXR and H108's radiologists with assistance

Findings	Rater 1 *vs.* AI		Rater 2 *vs.* AI		Rater 3 *vs.* AI	
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ
Lung Tumor	.950	.519 (.259, .780)	.885	.329 (.129, .530)	.950	.558 (.313, .802)
Pneumonia	.925	.776 (.668, .884)	.905	.735 (.623, .846)	.910	.703 (.575, .830)
Tuberculosis	.970	.807 (.657, .957)	.950	.723 (.560, .885)	.940	.707 (.553, .862)
Other Diseases	.735	.414 (.291, .537)	.700	.356 (.234, .477)	.815	.543 (.415, .670)
No finding	.910	.820 (.742, .898)	.900	.800 (.718, .882)	.935	.870 (.802, .938)
Aortic Enlargement	.880	.359 (.174, .544)	.900	.408 (.209, .606)	.915	.417 (.199, .635)
Atelectasis	.910	.091 (-.076, .259)	.905	.374 (.151, .597)	.905	.146 (-.057, .349)
Calcification	.905	.550 (.376, .723)	.865	.449 (.286, .612)	.870	.002 (-.138, .142)
Cardiomegaly	.925	.625 (.451, .799)	.920	.634 (.470, .799)	.935	.662 (.491, .832)
Consolidation	.925	.506 (.289, .722)	.905	.146 (-.057, .349)	.940	.594 (.386, .801)
ILD	.840	.518 (.372, .664)	.845	.537 (.394, .680)	.850	.449 (.286, .611)
Infiltration	.845	.304 (.134, .475)	.905	.712 (.592, .833)	.890	.627 (.485, .769)
Lung Opacity	.860	.338 (.147, .528)	.860	.255 (.058, .453)	.905	.308 (.087, .529)
Nodule/Mass	.875	.489 (.316, .661)	.880	.516 (.347, .685)	.880	.487 (.311, .663)
Pleural Effusion	.985	.954 (.901, 1.000)	.945	.839 (.747, .931)	.930	.773 (.659, .886)
Pleural Thickening	.915	.407 (.178, .636)	.855	.420 (.253, .587)	.910	.451 (.237, .664)
Pneumothorax	.990	.745 (.405, 1.000)	.985	.565 (.125, 1.000)	.985	.565 (.125, 1.000)
Pulmonary Fibrosis	.905	.587 (.420, .754)	.870	.508 (.347, .670)	.900	.599 (.441, .756)
Other Lesions	.945	.563 (.332, .794)	.955	.643 (.426, .859)	.935	.546 (.328, .764)
Mean	**.905**	**.546 (.370, .717)**	**.891**	**.524 (.348, .699)**	**.911**	**.527 (.342, .711)**

Impact of VinDr-CXR on radiologist diagnostic agreement

We show evidence that the VinDr-CXR helped improve the degree of agreement among radiologists for the task of detection of lung lesions (see Figure 5**A**). Specifically, Fleiss' Kappa values improved by 1.4% and 1.6% for HMUH and H108 readers, respectively. The inter-rater reliability has improved by 1.0%–2.9%, except for Reader 1 and Reader 2 from the HMUH (Δ = -0.2%). Additionally, we found that the rate of VinDr-CXR agreement with the participating radiologists was slightly higher than the rate of agreement among radiologists. With the support of VinDr-CXR, the agreement between the proposed DL system and radiologists significantly improved. As shown in Figure 5**B**, increments of agreement degree ranged 1.8%–6.5% after consulting the VinDr-CXR predictions. Figure 9 in the Supplementary Materials (p 19) shows several representative images from our reader study, which indicates the change in diagnostic decision after viewing the VinDr-CXR recommendation.

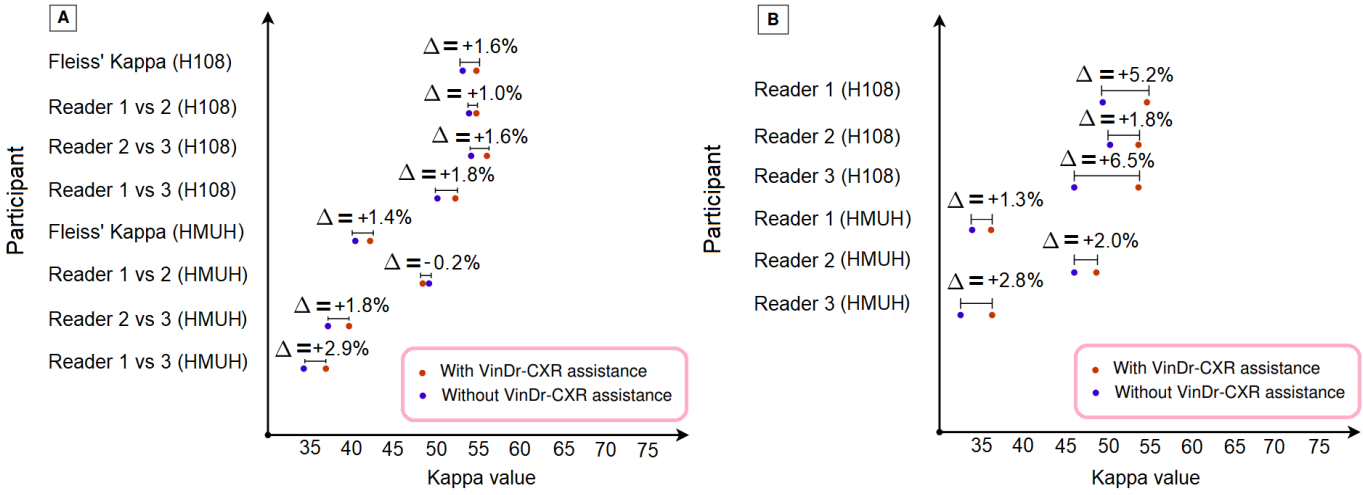

Figure 5. Impact of the VinDr-CXR on radiologists' performance. (**A**) Change in inter-radiologist agreement before and after consulting the VinDr-CXR predictions. The VinDr-CXR assistance significantly improved agreement between radiologists with an increase of 1.5% in mean Fleiss' Kappa. (**B**) Change in individual radiologists' agreement with the system before and after consulting the VinDr-CXR predictions. The VinDr-CXR assistance resulted in a significant increase in the agreement between the AI system and radiologists with an improvement of 3.3% in mean Cohen's Kappa. In particular, all differences were statistically significant.

Visual interpretability of VinDr-CXR

Explainability is an absolute necessity for the broad deployment of AI models in clinical practice. In a recent publication, the United States Food and Drug Administration (FDA) mentioned that explainability is required[36] for any AI-based computer-aided diagnosis (CAD) system. In summary, an interpretable AI/DL system can show the links between the features extracted by the system and its predictions[37]. Particularly, those links can be understood by a human expert[38]. Explainable DL systems

help human experts understand the underlying reasoning of DL systems and identify individual cases in which an AI model potentially gives incorrect predictions. In this study, the proposed VinDr-CXR is not only able to provide disease conclusions (global labels), but a helpful explanation involves abnormal findings (local labels) with their corresponding exact locations. Beyond the classification output, the VinDr-CXR can provide localization information that locates abnormalities accurately on CXR scans. The bounding boxes provided by the DL system may be an essential consideration supporting classification outputs. To illustrate the interpretability of the system, we utilized the trained classification network to compute and visualize the saliency maps for several examples from the internal validation set. To this end, we extracted feature maps produced by the last convolutional layer of the VinDr-CXR classifier model. We then used principal component analysis (PCA)[39] to reduce the channels of the feature map into a single channel. Then, this single-channel map is converted to a saliency map for visualization. Figure 6 shows the original CXRs and lesion bounding boxes annotated by our expert radiologists for Lung Tumor and Tuberculosis (TB) patients. The corresponding saliency maps obtained from the VinDr-CXR system are also provided. We observed the following insights that help better understand the decision-making process of the VinDr-CXR system. First, the visualization of normal cases was irregular with a symmetric high colormap, and there was no increased signal over all parts of the lung (Figure 6**A**). Second, across all abnormal scans, the saliency maps highlighted parts of the CXRs that contain abnormal patterns such as Nodule/Mass, Calcification, and Opacity, which are clinically correlated with disease conclusions, including Lung Tumor and TB. In other words, the attention regions in the visualization maps were consistent with the annotated abnormal findings provided by our radiologists, as well as predictions by the VinDr-CXR detector model (Figure 6**B–F**).

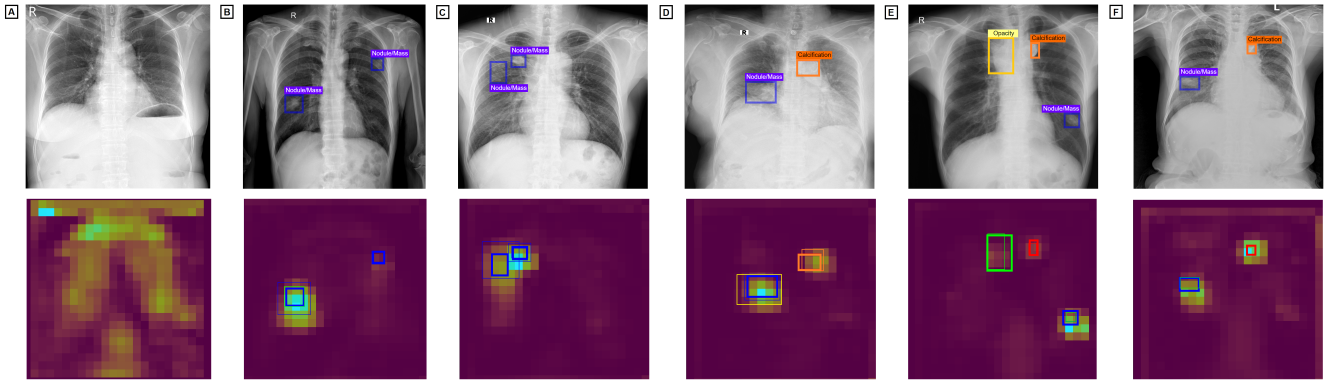

Figure 6. Visual explainability of the VinDr-CXR system. Visualization of CXRs and saliency maps overlaid on the original CXR images for several patients chosen from the internal validation set. The CXRs and affected areas were provided by our expert radiologist (top row). The corresponding saliency maps for CXRs were found by the VinDr-CXR classification model (bottom row). The blue, red, and green boxes cover the regions of Nodule/Mass, Calcification, and Opacity, respectively. Boxes with thick and thin lines denote ground-truth boxes and detected boxes, respectively. This figure indicates that the saliency maps highlighted parts of the CXRs that contain abnormal patterns (local labels), which are clinically correlated with disease conclusions (global labels). Best viewed in color.

Discussion

The VinDr-CXR may find helpful in several clinical scenarios. (1) The system was able to discriminate correctly between normal and abnormal CXRs. It, therefore, could be used as a tool to automate screening of common lung diseases for primary diagnosis (e.g., Tuberculosis and Malignancies) at scale. (2) Non-specialist clinicians could also potentially use the VinDr-CXR at healthcare centers to provide teleradiology primary CXR reading or to support rapidly triage cases. (3) The VinDr can be used as a second-reader or assistance tool for clinicians. In this setting, the system can provide diagnosis assistance at different levels including image-level diagnosis and lesion-level diagnosis. This makes the proposed DL algorithms more transparent and more explainable, allowing clinicians to understand and explore the results to improve diagnosis outcomes.

On the technical aspect, this study strongly supports the following statements. First, a DL network can learn effectively and accurately predict common thoracic diseases and findings if trained on large-scale, multi-institutional, and expert-annotated imaging datasets. Second, although the current study was conducted in Vietnam, the proposed VinDr-CXR showed its robustness on external datasets acquired from hospitals in the United States. This result showed evidence that the model generalization of a DL system could consistent across geographical settings. Several studies[40,41] showed that DL models for

the CXR interpretation failed to generalize to image sources from new institutions and hospitals. In contrast, our finding is similar to Cohen *et al.*[42] who presented evidence that generalization over difference distributions is not due to a shift in the images but instead a shift in the labels. Third, our external validation experiment showed evidence that a DL system trained on digital CXR could generalize well on CXR captured by smartphone cameras without additional training cost. This opens the opportunity to integrate the DL system like VinDr-CXR into the vast spectrum of clinical workflows across the world, including developing regions that is still using films.

In terms of novelty, the VinDr-CXR system shows several contributions. First, while most previous studies[14–19] used machine-generated annotations that contain many CXR images with uncertainty labels to train DL systems, our model was trained on radiologist-generated annotations for both development and validation data sets. Next, the proposed DL system evaluated both the detection and classification tasks, while most previous studies have only evaluated image-level classification performance without specifying the location of abnormal findings. Last, this study conducted a large-scale, clinical evaluation to investigate the actual impact of a DL system on the variability in radiologist performance in the interpretation of CXRs. We showed that the system significantly improved agreement among radiologists. To the best of our knowledge, we are the first to show that a DL system trained on a large-scale, annotated dataset can offer clinical value by helping to improve the rate of agreement among physicians. Furthermore, we also observed that VinDr-CXR assistance resulted in a significant increase in the agreement between the DL system and radiologists. Note that most literature refers to comparisons between human performing and AI models on the CXR interpretation, usually in diagnostic accuracy on the same clinical validation dataset. We suggest that these comparisons do not offer valuable insights into these systems' impact on clinical practice.

This study is not without limitations. First, the development and evaluation datasets only contain frontal CXR scans. Meanwhile, several clinical findings require lateral views. The next version of the dataset may consider adding the lateral views to train the VinDr-CXR system. Second, in clinical practice, physicians diagnose diseases based on both the patient's clinical history and visual information from CXRs. The VinDr-CXR, however, used only image information for providing diagnosis results without taking clinical and laboratory information into account. Third, the current study's most significant limitation is that we did not directly measure the actual impact of the VinDr-CXR on the sensitivity and specificity of participating radiologists due to the lack of gold reference ground truth. We showed in this study that the DL system was able to reduce clinical disagreements among radiologists. However, there is no clear evidence that the DLS helps improve the sensitivity or specificity of the radiologist in CXR interpretation.

In conclusion, a DL system, namely VinDr-CXR, for classifying and localizing the common thoracic diseases and crucial findings was developed and externally validated in this study. The system showed its high diagnostic accuracy across internal and external populations. Our results support the assertion that DL models can improve current clinical practice by improving agreement among clinical experts during medical imaging interpretation. The proposed system could be directly applied in different clinical settings, e.g., supporting physicians in triaging cases or using as a second reader. Although many DL–based models for predicting lung diseases have improved diagnostic accuracy, in some cases surpassing radiologists' performance, there is little evidence showing that deployment of these models has improved patient outcomes. Therefore, further research is needed to validate the model prospectively and determine its utility in clinical settings. For example, the diagnostic and clinical effects of the VinDr-CXR needs to be assessed in large-scale test cohorts to determine the change in sensitivity and specificity of radiologists for the CXR interpretation in routine clinical practice.

Methods

Our study was approved by the Institutional Review Boards (IRB) of the HMUH and H108. In addition, the requirement for obtaining informed patient consent was waived due to the observational nature of this study. Under these approvals, raw CXR images in DICOM standard were collected retrospectively. Protected health information (PHI) has been de-identified to comply with the regulations of the U.S. HIPAA[43], European GDPR[44], and the local privacy laws[45]. In this section, we describe the methodology for developing and validating the VinDr-CXR system. First, we provide details of datasets used in this study. Next, the development of DL algorithms is introduced. Last, we describe the experimental design for our reader study and statistical analysis methods used for model evaluation.

Datasets for VinDr-CXR development and validation

VinDr-CXR dataset

To develop and validate the VinDr-CXR, we retrospectively collected a total of more than 100,000 anterior–posterior (AP) and posterior–anterior (PA) CXR scans of adult patients (aged > 10 years). The imaging data were in DICOM format and performed at two major hospitals (*i.e.*, the HMUH and H108) between January 1, 2018 and December 31, 2020. In addition, CXR studies were acquired from a wide diversity of scanners and manufacturers such as Phillips, GE, Fujifilm, Siemens, Toshiba, Canon, and Samsung. Out-of-distribution samples (e.g., images with poor quality or invalid) were excluded via a manual inspection process. A group of 17 expert radiologists has labeled a portion of the raw dataset, and each has at least ten years of experience. Overall, the dataset contains 54,485 CXR studies (mean age 43.77 years; 47.79% female patients) that meet the study criteria. A development set consists of 51,485 studies used to optimize DL algorithms and an internal validation set of 3,000 studies to evaluate models' performance. All CXRs were performed in independent patients for the validation set and did not overlap with the development set to avoid bias. Each CXR scan in the development set was labeled by three independent radiologists, while a panel of five board-certified radiologists labeled each case in the internal validation set, and their consensus established the reference standard. Ground truth was established at the image-level for the classification task and pixel-level for the localization task. Our dataset was labeled for the presence of 28 labels. We defined for the first time two different types of labels in CXRs: (1) global labels that are image-level labels representing diseases or impressions and (2) local labels (lesion-level annotations) that are critical findings or lesions that occur in CXRs. The participating radiologists provide precisely the location of lesions or abnormalities via bounding box annotations for local labels. The data collection and annotation process is summarized in Figure 1B. The cohort demographic information and statistics of development and internal validation data sets are summarized in Table 8. Full details of the dataset collection and labeling process can be found in our previous study[46].

External datasets

We used two public CXR datasets to assess the accuracy and efficiency of the VinDr-CXR system across populations. Our external validation tests used data from patients in the CheXpert[16] (*N = 200*) and CheXphoto[33] (*N = 200*) datasets. These datasets shared several disease labels with the VinDr-CXR dataset, such as Pleural Effusion, Pneumonia, and No Finding. The CheXpert[16] is a large public dataset for the CXR interpretation performed between 2002 and 2017 at Stanford Hospital, USA. The CheXpert validation set contains 200 studies, for which the ground-truth label of each study is obtained by taking the majority vote of three board-certified radiologists. The CheXphoto[33] is a recently published CXR dataset for the automated interpretation of photos of CXR through cell phone photography. The CheXphoto validation set comprises natural photos of all 200 studies in the CheXpert validation set. It can be used as a resource for testing the robustness of DL algorithms on smartphone photos of CXRs. Additional details of the external datasets are provided in Irvin *et al*.[16] and Phillips *et al*.[33].

Model development and training

To develop the VinDr-CXR system, we used a total of 51,485 annotated CXR scans from the development set. The system takes a CXR as input, and outputs are both disease classification and lesion localization. The whole VinDr-CXR architecture consists of two subnetworks, including a classification network and a detection network (see Figure 1A). We trained the EfficientNet-B6[31] model on the CXR images with a size of 1024×1024 pixels to classify common chest diseases. The EfficientNet[31] was well-known as a state-of-the-art DL architecture for image recognition tasks. It can achieve a high level of accuracy while requiring less computational cost for model training. We used mean binary cross-entropy loss to optimize the network in a supervised manner using image-level annotations. To localize abnormal findings, we deployed EfficientDet-D6[32], a recent advance of DL-based detector for the object detection tasks. The per-lesion annotations provided by radiologists were used to optimize the EfficientDet-D6[32] network. To reduce the impact of class imbalance, we adopted the focal loss[47] to optimize the detection network's weights. Several data augmentation strategies have been applied to minimize the risk of over-fitting in both two networks. Both the classification and detection networks were implemented using Python 3.7 (`https://www.python.org/`), PyTorch 1.6 (`https://pytorch.org/`), and trained on an NVIDIA V100 32GB GPU. A detailed description of the model development and training is provided in the Supplementary Materials, pp 16–17.

Reader study

To validate the effectiveness of the proposed DL approach, we conducted a reader study to assess the actual impact of the VinDr-CXR on the agreement of participating radiologists. We describe the reader study as bellows.

Data collection. For clinical evaluation, 400 CXR examinations were collected retrospectively from the HMUH and H108 under IRB approvals. These examinations were acquired between March 2021 and June 2021 after the training process of the VinDr-CXR has been completed. Among 400 CXR studies, half (*N = 200*) was obtained from the HMUH, and the rest (*N = 200*) was from the H108. Data sampling was conducted based on the actual distributions at the hospitals. This ensures that the imaging data are representative of the real-world conditions in which the DL algorithms will be deployed. In addition, the CXR

Table 8. Characteristics of the datasets used for VinDr-CXR development and internal validation

	Characteristics	Development set	Internal validation set
Collection statistics	Years	2018 to 2020	2018 to 2020
	Number of studies	51,485	3,000
	Number of images	51,941	3,000
	Number of abnormal studies	26,887	949
	Number of normal studies	24,598	2,051
	Number of human annotators per scan	3	5
	Image size (pixel×pixel, mean)	2,729 × 2,395	2,748 × 2,394
	Age (years, mean)*	51.28	31.80
	Male (%)*	52.77	55.90
	Female (%)*	47.23	44.10
	Data size (GiB)	509.5	31.3
Local labels	1. Aortic Enlargement	9,183 (17.68%)	221 (07.37%)
	2. Atelectasis	1,346 (2.59%)	96 (3.20%)
	3. Cardiomegaly	8,359 (16.09%)	310 (10.33%)
	4. Calcification	3,154 (6.07%)	232 (7.73%)
	5. Clavicle Fracture	300 (0.58%)	2 (0.07%)
	6. Consolidation	2,125 (4.09%)	126 (4.20%)
	7. Edema	61 (0.12%)	0 (0%)
	8. Emphysema	1,007 (1.94%)	4 (0.13%)
	9. Enlarged PA	451 (0.87%)	9 (0.30%)
	10. Interstitial Lung Disease (ILD)	5,946 (11.45%)	316 (10.53%)
	11. Infiltration	3,254 (6.26%)	79 (2.63%)
	12. Lung Cavity	175 (0.34%)	10 (0.33%)
	13. Lung Cyst	129 (0.25%)	3 (0.10%)
	14. Lung Opacity	4,680 (9.01%)	95 (3.17%)
	15. Mediastinal Shift	706 (1.36%)	20 (0.67%)
	16. Nodule/Mass	5,521 (10.63%)	286 (9.53%)
	17. Pulmonary Fibrosis	6,919 (13.32%)	358 (11.93%)
	18. Pneumothorax	321 (0.62%)	25 (0.83%)
	19. Pleural Thickening	7,899 (15.21%)	240 (8.00%)
	20. Pleural Effusion	4,554 (8.76%)	136 (4.53%)
	21. Rib Fracture	1,241 (2.39%)	17 (0.57%)
	22. Other Lesions	3,851 (7.41%)	112 (3.73%)
Global labels	23. Lung Tumor	1,650 (3.18%)	80 (2.67%)
	24. Pneumonia	3,827 (7.37%)	246 (8.20%)
	25. Tuberculosis	2,130 (4.10%)	164 (5.47%)
	26. Other Diseases	18,848 (36.29%)	657 (21.90%)
	27. COPD	388 (0.75%)	2 (0.07%)
	28. No Finding	16,461 (31.7%)	2,051 (68.37%)

(*) The calculations were performed based on the number of CXR scans for which sex and age were known. For global labels, the number of positive examples was reported based on the majority vote of binarized labels provided by radiologists. For local labels, the percentage rate of number bounding boxes on the number of studies was reported. Several abnormalities (Clavicle Fracture, Edema, Emphysema, Enlarged PA, Lung Cavity, Lung Cyst, Mediastinal Shift, Rib Fracture, and COPD) were not considered during model training because the number of positive examples in the internal validation set was very limited. The “No Finding” label was intended to capture the absence of all findings and pathologies.

scans will be used to evaluate the agreement among participating readers. Hence, we did not establish a reference standard for the collected data.

Reader selection. We recruited a group of six board-certified radiologists from the radiology departments of the HMUH and H108 to participate in our observer performance test. All participating radiologists were trained in CXR interpretation and had an average of 15.5 years of clinical experience interpreting thoracic diseases (range 10–22 years). In addition, the readers read an average of 25,000 CXR scans each year (range 15,000–40,000). Table 9 shows the characteristics of radiologists who participated in our reader study.

Table 9. Characteristics of the participating radiologists. Mean annual diagnostic volumes were calculated based on the number of CXR scans interpreting.

Reader	Year's experience	Mean annual diagnostic volume (studies)
Radiologist 1 (H108)	10	40,000
Radiologist 2 (H108)	15	14,000
Radiologist 3 (H108)	11	45,000
Radiologist 1 (HMUH)	21	20,000
Radiologist 2 (HMUH)	22	15,000
Radiologist 3 (HMUH)	14	15,000
Average	**15.5**	**25,000**

Reader study design. The reader study was conducted in two sessions. In the first session, participating readers read the CXR scans independently without the VinDr-CXR assistance. During the second session, the readers re-evaluated all CXR scans with the assistance of the VinDr-CXR. Specifically, the radiologists were provided the VinDr-CXR predictions in the form of bounding boxes, which locate abnormalities (see Figure 10; Supplementary Materials, p 20). They considered the model's prediction and modified the diagnostics. During this process, the readers were blinded to the relevant clinical information such as the original reports and previous medical histories of the patients or other patient records. To maximize human performance, the readers can perform the task on our browser-based viewer with zoom in or out, panning, and many other support tools. The reader study was set up to ensure that all radiologists can view and interpret the CXR studies in an environment similar to their routine workflow in clinical practice. Changes in radiologists' agreement were then assessed to investigate the impact of the VinDr-CXR assistance.

Statistical analysis

Diagnostic performance metrics, including area under the receiver operating characteristic curve (AUC), sensitivity, specificity, *F1*-score, false-positive rate (FPR), and false-negative rate (FNR), were used to assess the accuracy of the VinDr-CXR for the classification task. For each indicator, 95% confidence interval (CI) was estimated with bootstrapping (10,000 replications). To evaluate the VinDr-CXR's ability to detect and localize lung lesions, we used the FROC (the sensitivities of models under different false positive rates as 0.25, 0.5, 1, 2, and 4). Cohen's Kappa statistics[48] and percentage agreement rate were used to evaluate the level of agreement between the VinDr-CXR system and participating radiologists, as well as to assess the agreement between pairs of radiologists. To assess inter-rater agreement among a group of radiologists, the Fleiss' Kappa[49] score was used. The Kappa values were interpreted as following guidelines[50]: (< 0.00): poor; (0.00–0.20): none to slight; (0.21–0.40): fair; (0.41–0.60): moderate; (0.61–0.80): substantial; and (0.81–1.00): almost perfect agreement. All statistical analyses were performed using Python (version 3.9.2 – https://www.python.org/) and scikit-learn (version 0.24.2 – https://scikit-learn.org/).

Data availability

To facilitate a wide range of research topics in computer vision and medical imaging, we made the VinDr-CXR dataset (18,000 studies) publicly available through through PhysioNet at https://physionet.org/content/vindr-cxr/1.0.0/. The image and annotation quality of the dataset can be visually check via our project webpage at https://vindr.ai/datasets/cxr. The CheXpert dataset is publicly available at https://stanfordmlgroup.github.io/competitions/chexpert/. The CheXphoto dataset is publicly available at https://stanfordmlgroup.github.io/competitions/chexphoto/.

Code availability

Implementation of our work is based on the following open source repositories: Pytorch: https://pytorch.org; OpenCV https://opencv.org/; Pydicom: https://pydicom.github.io/. The source code used to train the VinDr-CXR system is a part of a commercial software product and not available to the public. The commercial version of VinDr-CXR

can be freely tried through an online demonstration at https://vindr.ai/. All performance metrics were calculated with the support of scikit-learn https://scikit-learn.org/. Our labeling framework called VinDr Lab was made as open-source software and available for downloading at https://vindr.ai/vindr-lab.

Acknowledgements

This research was supported by Vingroup Big Data Institute. The authors would like to acknowledge the Hanoi Medical University Hospital (HMUH) and Hospital 108 (H108), for providing us access to their databases and agreeing to make the VinDr-CXR dataset publicly available. We are especially thankful to all the radiologists participating in the reader study.

Author contributions

H.Q.N., M.D., and V.V. directed the project; H.H.P. and H.Q.N. contributed to the conception and design of the study; H.Q.N., K.L., and L.T.L. carried out the prospective experiments; H.H.P., D.B.N., H.T.N., T.T.L., and T.V.N. contributed to the analysis of the data and developed the deep learning models; H.H.P. wrote the manuscript with the assistance and feedback of all the other co-authors.

Competing interests

This study was funded by the Vingroup Big Data Institute (VinBigdata). H.Q.N., H.H.P., M.D., D.B.N., H.T.N., T.T.L., and T.V.N. were employed by VinBigdata. All other authors declare no competing interests.

Supplementary Materials

Development of deep learning algorithms

Network architectures. Deep learning (DL)[51], a subfield of machine learning, is a computational model that composes multiple processing layers and uses data-driven rules to learn representations of data with multiple levels of abstraction. DL networks showed their breakthrough successes in a wide variety of diagnostic tasks in medical imaging analysis[13,14,23,52]. In this study, we applied two well-known DL networks for classifying common thoracic diseases and detecting abnormal findings in CXR images. We deployed EfficientNet-B6[31] for the task of disease classification and EfficientDet-D6[32] for the task of lesion detection. These network architectures were well-known as the most commonly used and most successful DL networks for image classification and object detection. By balancing network depth, width, and resolution, the EfficientNet[31] can lead to much better accuracy and efficiency than other state-of-the-art CNN models. Meanwhile, the EfficientDet-D6[32] used a weighted bi-directional feature pyramid network (BiFPN), allowing easy and fast multiscale feature fusion and combined with the feature learning capacity of EfficientNet-B6[31]. This network architecture design, which requires less computational resources for training, achieved much better efficiency than prior state-of-the-art detectors. Therefore, we found that EfficientNet-B6[31] and EfficientDet-D6[32] well-suited for medical applications, including the CXR analysis. Figure 7 illustrates the key ideas behind these two network architectures.

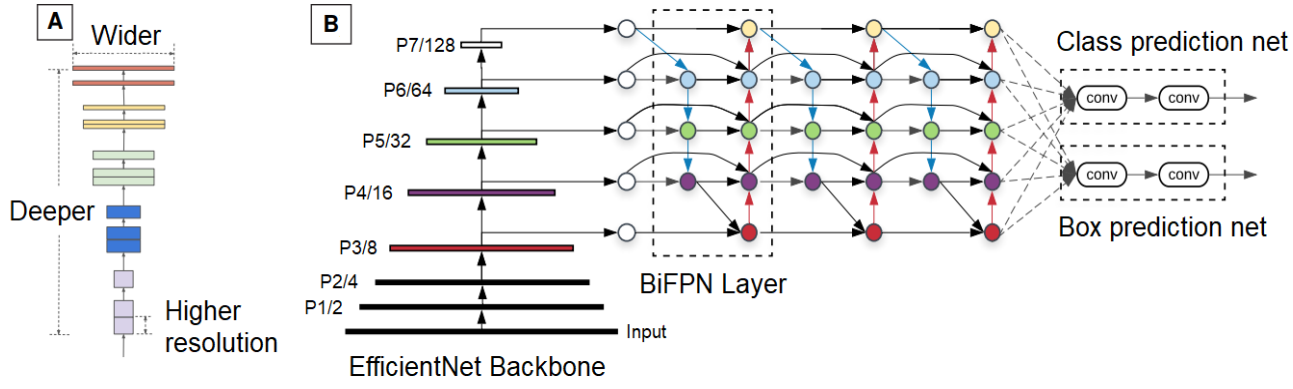

Figure 7. (**A**) EfficientNet architecture: scaling method that uniformly scales all three dimensions with a fixed ratio. (**B**) EfficientDet architecture employs EfficientNet as the backbone network, BiFPN as the feature network, and shared class/box prediction network. These figures were reproduced from the original papers.

Implementation details. We followed the original implementation of EfficientNet-B6[31] and EfficientDet-D6[32] with some minor modifications. Specifically, the last fully-connected layer of EfficientNet-B6[31] has changed to output a vector of six dimensions corresponding to six classes. Mean binary cross-entropy loss was used to train these six classes simultaneously instead of a single multi-class cross-entropy loss in the original works. For the detector, we only changed the classification head to return scores for 14 lesion types while the loss functions and other parts of the architecture were preserved.

Data augmentation. Several pre-processing steps were performed on CXRs before passing to the networks. Images in DICOM format were converted to 8-bit PNG, then padded and resized to 1024×1024 pixels. Subsequently, 1-channel images were transformed to 3-channel ones by repeating the channel three times. In the last step, images with the pixels intensity in range $[0, 255]$ were normalized by subtracting $(123.675, 116.28, 103.53)$ then dividing by $(58.395, 57.12, 57.375)$ in a channel-wise manner.

Training procedures. Regarding the training procedure, the classifier's weights were initialized with weights trained on the ImageNet dataset[53], a large-scale natural image dataset for the classification task. Pre-processed training images were randomly transformed using resize-cropping, shift-scale-rotating, horizontal flipping, brightness-contrast adjustments, then grouped into batches of 16. The classifier was trained for multi-label binary classification tasks (presence or absence of each disease) by optimizing the mean binary cross-entropy loss of all diseases. A variant of the stochastic gradient descent algorithm, Adam optimizer[54], was used with a base learning rate of $2 \times 10e - 4$. The learning rate was then linearly increased in the first epoch then gradually annealed to 0 at the end of the 50-*th* epoch, following the cosine function. For the detector, a similar training method was employed. Detector's weights were initialized with parameters trained on the COCO dataset[55], a large dataset with common objects marked by bounding boxes. The optimization objective incorporates both regression loss for box categories by weighted summation. A batch size of 8 was used due to larger memory consumption compared to the classifier and the total training schedule was 60 epochs.

Clinical evaluation results at the HMUH

Table 10. Inter-rater agreement among radiologists at the HMUH without the VinDr-CXR assistance

Findings	Rater 1 *vs.* Rater 2		Rater 1 *vs.* Rater 3		Rater 2 *vs.* Rater 3		Fleiss' κ
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ	
Lung Tumor	.955	.617 (.388, .846)	.945	.591 (.372, .810)	.960	.646 (.419, .873)	.616 (.536, .696)
Pneumonia	.865	.536 (.388, .684)	.815	.312 (.157, .468)	.870	.281 (.080, .482)	.383 (.303, .463)
Tuberculosis	.925	.675 (.522, .827)	.915	.385 (.164, .605)	.880	.298 (.111, .486)	.474 (.394, .554)
Other Diseases	.775	.470 (.344, .596)	.660	.234 (.105, .363)	.735	.440 (.312, .567)	.373 (.293, .453)
No Finding	.880	.758 (.668, .848)	.865	.730 (.636, .825)	.905	.811 (.731, .890)	.766 (.686, .846)
Aortic Enlargement	.955	.549 (.289, .810)	.915	-.032 (-.056, -.007)	.940	-.027(-.047, -.008)	.207 (.127, .287)
Atelectasis	.980	.490 (.060, .920)	.975	-.008 (-.021, .005)	.975	-.008 (-.021, .005)	.210 (.130, .290)
Calcification	.925	.610 (.432, .789)	.920	.556 (.363, .749)	.945	.735 (.586, .884)	.638 (.558, .718)
Cardiomegaly	.890	.277 (.076, .478)	.925	.318 (.067, .568)	.885	.404 (.207, .601)	.333 (.253, .413)
Consolidation	.935	.216 (-.046, .478)	.955	.164 (-.144, .471)	.930	.330 (.066, .593)	.249 (.169, .329)
ILD	.920	.488 (.275, .702)	.815	.131 (-.028, .290)	.845	.352 (.175, .529)	.303 (.223, .383)
Infiltration	.885	.000 (.000, .000)	.980	.000 (.000, .000)	.885	.118 (-.056, .292)	.030 (-.050, .110)
Lung Opacity	.785	.229 (.072, .386)	.785	.103 (-.020, .225)	.870	.020 (-.114, .154)	.116 (.036, .196)
Nodule/Mass	.955	.688 (.498, .877)	.915	.475 (.267, .682)	.900	.490 (.298, .682)	.540 (.460, .620)
Pleural Effusion	.925	.509 (.298, .721)	.945	.674 (.494, .853)	.930	.425 (.178, .673)	.547 (.467, .627)
Pleural Thickening	.940	.377 (.106, .648)	.975	.275 (-.164, .714)	.925	.102 (-.100, .304)	.245 (.165, .325)
Pneumothorax	1.00	1.00 (1.00, 1.00)	1.00	1.00 (1.00, 1.00)	1.00	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)
Pulmonary Fibrosis	.905	.545 (.367, .724)	.895	.343 (.125, .560)	.870	.367 (.177, .557)	.423 (.343, .503)
Other Lesions	.890	.175 (-.024, .375)	.920	.165 (-.073, .402)	.890	.332 (.120, .544)	.232 (.152, .312)
Mean	**.910**	**.485 (.301, .669)**	**.901**	**.338 (.171, .504)**	**.902**	**.374 (.217, .532)**	**.404 (.329, .480)**

Table 11. Inter-rater agreement among radiologists at the HMUH with the VinDr-CXR assistance

Findings	Rater 1 *vs.* Rater 2		Rater 1 *vs.* Rater 3		Rater 2 *vs.* Rater 3		Fleiss' κ
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ	
Lung Tumor	.955	.617 (.388, .846)	.950	.640 (.433, .846)	.965	.702 (.494, .910)	.652 (.572, .732)
Pneumonia	.865	.536 (.388, .684)	.820	.338 (.181, .494)	.875	.323 (.121, .524)	.403 (.323, .483)
Tuberculosis	.925	.685 (.537, .833)	.915	.385 (.164, .605)	.870	.279 (.100, .459)	.471 (.391, .551)
Other Diseases	.775	.470 (.344, .596)	.665	.241 (.111, .371)	.740	.449 (.322, .576)	.379 (.299, .459)
No Finding	.880	.758 (.668, .848)	.870	.740 (.647, .833)	.900	.801 (.719, .882)	.766 (.686, .846)
Aortic Enlargement	.945	.565 (.337, .793)	.920	.243 (.001, .486)	.935	.103 (-.134, .341)	.340 (.260, .420)
Atelectasis	.975	.432 (.021, .842)	.970	-.008 (-.022, .005)	.975	-.008 (-.021, .005)	.186 (.106, .266)
Calcification	.930	.656 (.490, .823)	.940	.694 (.531, .856)	.940	.716 (.564, .868)	.689 (.609, .769)
Cardiomegaly	.885	.267 (.071, .463)	.930	.335 (.076, .593)	.885	.405 (.210, .601)	.333 (.253, .413)
Consolidation	.935	.216 (-.046, .478)	.955	.164 (-.144, .471)	.930	.330 (.066, .593)	.249 (.169, .329)
ILD	.930	.575 (.376, .773)	.815	.161 (-.002, .324)	.855	.412 (.238, .585)	.361 (.281, .441)
Infiltration	.880	.000 (.000, .000)	.980	.000 (.000, .000)	.880	.112 (-.055, .280)	.026 (-.054, .106)
Lung Opacity	.775	.212 (.057, .367)	.780	.099 (-.020, .219)	.865	.017 (-.113, .146)	.106 (.026, .186)
Nodule/Mass	.950	.662 (.469, .855)	.915	.500 (.299, .701)	.905	.525 (.338, .712)	.554 (.474, .634)
Pleural Effusion	.925	.509 (.298, .721)	.945	.674 (.494, .853)	.930	.425 (.178, .673)	.547 (.467, .627)
Pleural Thickening	.930	.338 (.084, .593)	.975	.275 (-.164, .714)	.915	.089 (-.092, .270)	.219 (.139, .299)
Pneumothorax	1.00	1.00 (1.00, 1.00)	1.00	1.00 (1.00, 1.00)	1.000	1.000 (1.00, 1.00)	1.00 (1.00, 1.00)
Pulmonary fibrosis	.890	.484 (.301, .666)	.895	.343 (.125, .560)	.875	.403 (.216, .590)	.413 (.333, .493)
Other Lesions	.895	.189 (-.013, .392)	.915	.158 (-.066, .383)	.890	.362 (.153, .571)	.248 (.168, .328)
Mean	**.908**	**.483 (.304, .662)**	**.903**	**.367 (.192, .543)**	**.902**	**.392 (.227, .557)**	**.418 (.342, .494)**

Table 12. Agreement between the VinDr-CXR system and HMUH's radiologists without assistance

Findings	Rater 1 *vs.* AI		Rater 2 *vs.* AI		Rater 3 *vs.* AI	
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ
Lung Tumor	.940	.427 (.164, .690)	.975	.693 (.439, .947)	.945	.451 (.182, .719)
Pneumonia	.845	.513 (.366, .661)	.890	.559 (.396, .721)	.870	.417 (.239, .594)
Tuberculosis	.910	.628 (.472, .784)	.945	.793 (.675, .910)	.865	.271 (.096, .446)
Other Diseases	.765	.390 (.248, .532)	.840	.640 (.528, .752)	.685	.314 (.181, .447)
No finding	.860	.721 (.625, .816)	.870	.742 (.652, .833)	.845	.689 (.589, .790)
Aortic Enlargement	.945	.616 (.409, .823)	.930	.430 (.188, .672)	.910	.154 (-.058, .366)
Atelectasis	.970	.235 (-.167, .636)	.980	.490 (.060, .920)	.985	.395 (-.146, .936)
Calcification	.900	.237 (.011, .462)	.895	.354 (.146, .562)	.920	.463 (.246, .680)
Cardiomegaly	.875	.203 (.017, .389)	.925	.674 (.520, .828)	.870	.340 (.146, .535)
Consolidation	.960	.318 (-.026, .662)	.925	.310 (.052, .568)	.935	.201 (-.073, .476)
ILD	.850	.178 (-.006, .361)	.880	.433 (.245, .621)	.815	.286 (.113, .459)
Infiltration	.885	.000 (.000, .000)	.880	.410 (.217, .604)	.885	.118 (-.056, .292)
Lung Opacity	.785	.179 (.030, .329)	.880	.269 (.060, .479)	.900	.045 (-.124, .215)
Nodule/Mass	.880	.378 (.190, .566)	.875	.442 (.261, .623)	.855	.390 (.211, .570)
Pleural Effusion	.945	.640 (.447, .833)	.940	.423 (.154, .691)	.970	.754 (.565, .943)
Pleural Thickening	.945	-.028 (-.045, -.011)	.895	-.045 (-.072, -.018)	.970	.239 (-.158, .635)
Pneumothorax	.990	.496 (-.104, 1.00)	.990	.496 (-.104, 1.00)	.990	.496 (-.104, 1.00)
Pulmonary Fibrosis	.880	.302 (.093, .511)	.865	.375 (.188, .562)	.865	.196 (-.003, .396)
Other Lesions	.935	-.033 (-.052, -.015)	.895	.235 (.023, .447)	.905	.051 (-.126, .228)
Mean	**.898**	**.337 (.141, .528)**	**.909**	**.459 (.244, .670)**	**.894**	**.330 (.090, .565)**

Table 13. Agreement between the VinDr-CXR system and HMUH's radiologists with assistance

Findings	Rater 1 *vs.* AI		Rater 2 *vs.* AI		Rater 3 *vs.* AI	
	Agreement	Cohen's κ	Agreement	Cohen's κ	Agreement	Cohen's κ
Lung Tumor	.940	.427 (.164, .690)	.975	.693 (.439, .947)	.950	.523 (.267, .778)
Pneumonia	.845	.513 (.366, .661)	.890	.559 (.396, .721)	.875	.447 (.271, .623)
Tuberculosis	.910	.628 (.472, .784)	.945	.798 (.683, .913)	.865	.271 (.096, .446)
Other Diseases	.765	.390 (.248, .532)	.840	.640 (.528, .752)	.690	.322 (.188, .455)
No Finding	.860	.721 (.625, .816)	.870	.742 (.652, .833)	.850	.700 (.601, .799)
Aortic Enlargement	.955	.718 (.544, .893)	.940	.542 (.315, .769)	.915	.231 (-.003, .464)
Atelectasis	.965	.205 (-.162, .571)	.980	.490 (.060, .920)	.985	.395 (-.146, .936)
Calcification	.905	.343 (.119, .566)	.895	.354 (.146, .562)	.925	.511 (.302, .721)
Cardiomegaly	.875	.203 (.017, .389)	.940	.743 (.605, .881)	.875	.354 (.158, .550)
Consolidation	.960	.318 (-.026, .662)	.925	.310 (.052, .568)	.935	.201 (-.073, .476)
ILD	.865	.275 (.080, .469)	.885	.466 (.281, .651)	.820	.315 (.143, .487)
Infiltration	.885	.000 (.000, .000)	.885	.446 (.255, .636)	.885	.118 (-.056, .292)
Lung Opacity	.780	.174 (.027, .321)	.885	.318 (.108, .527)	.900	.045 (-.124, .215)
Nodule/Mass	.875	.364 (.176, .551)	.875	.442 (.261, .623)	.870	.461 (.286, .636)
Pleural Effusion	.945	.640 (.447, .833)	.940	.423 (.154, .691)	.970	.754 (.565, .943)
Pleural Thickening	.945	-.028 (-.045, -.011)	.895	.045 (-.116, .205)	.970	.239 (-.158, .635)
Pneumothorax	.990	.496 (-.104, 1.00)	.990	.496 (-.104, 1.00)	.990	.496 (-.104, 1.00)
Pulmonary Fibrosis	.880	.302 (.093, .511)	.860	.364 (.178, .549)	.865	.196 (-.003, .396)
Other Lesions	.940	-.030 (-.048, -.012)	.895	.235 (.023, .447)	.915	.223 (-.017, .463)
Mean	**.899**	**.350 (.158, .538)**	**.911**	**.479 (.259, .694)**	**.897**	**.358 (.115, .595)**

Examples of CXR scans from internal and external datasets

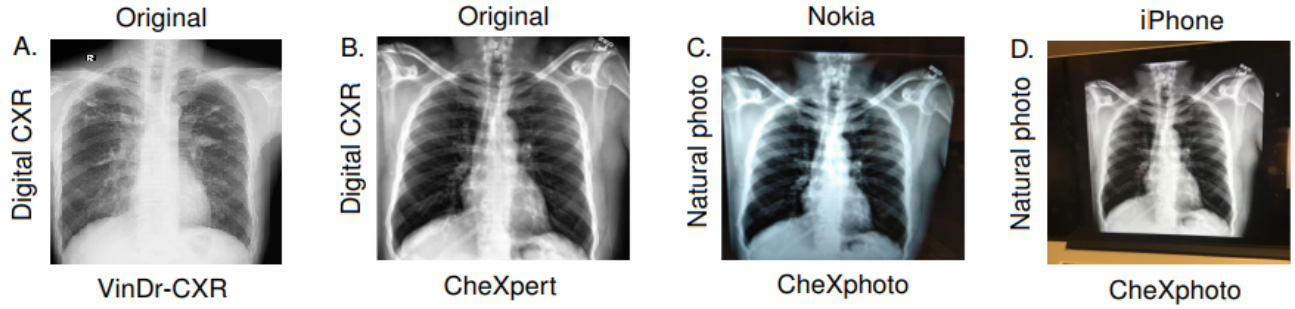

Figure 8. Examples of CXR scans from internal and external datasets. (A) An original CXR scan in DICOM format from the VinDr-CXR dataset. (B) An original CXR scan in Portable Graphics Format (.PNG) format from CheXpert[16] dataset. (C) A CXR image produced by a Nokia phone from CheXphoto[33] dataset. (D) A CXR image produced by iPhone from CheXphoto[33] dataset.

Representative cases from the reader study

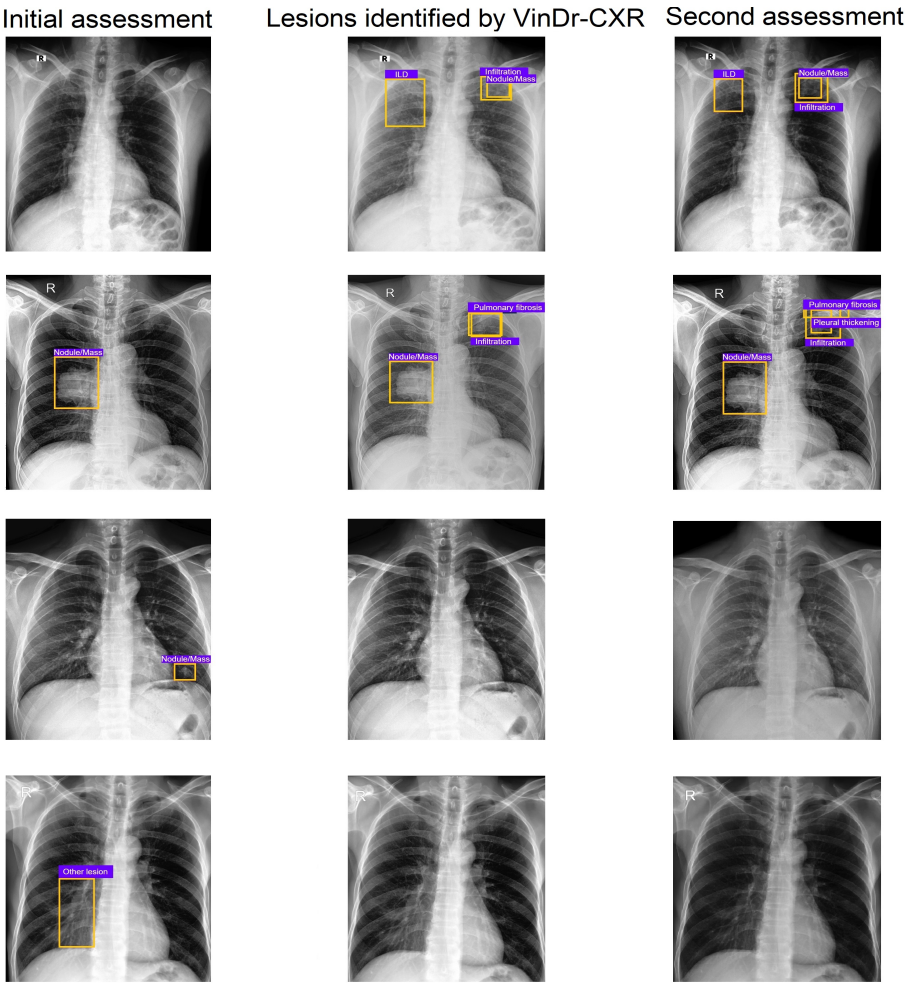

Figure 9. Representative cases from our reader study. The first column shows CXRs and lesions marked by participating radiologists for the first assessment. The middle column shows lesions identified by the VinDr-CXR system. The last column shows the final decision of the radiologists after consulting the VinDr-CXR's result. In many cases, radiologists have changed their previous decisions by adding or removing lesions.

VinDr-CXR interface

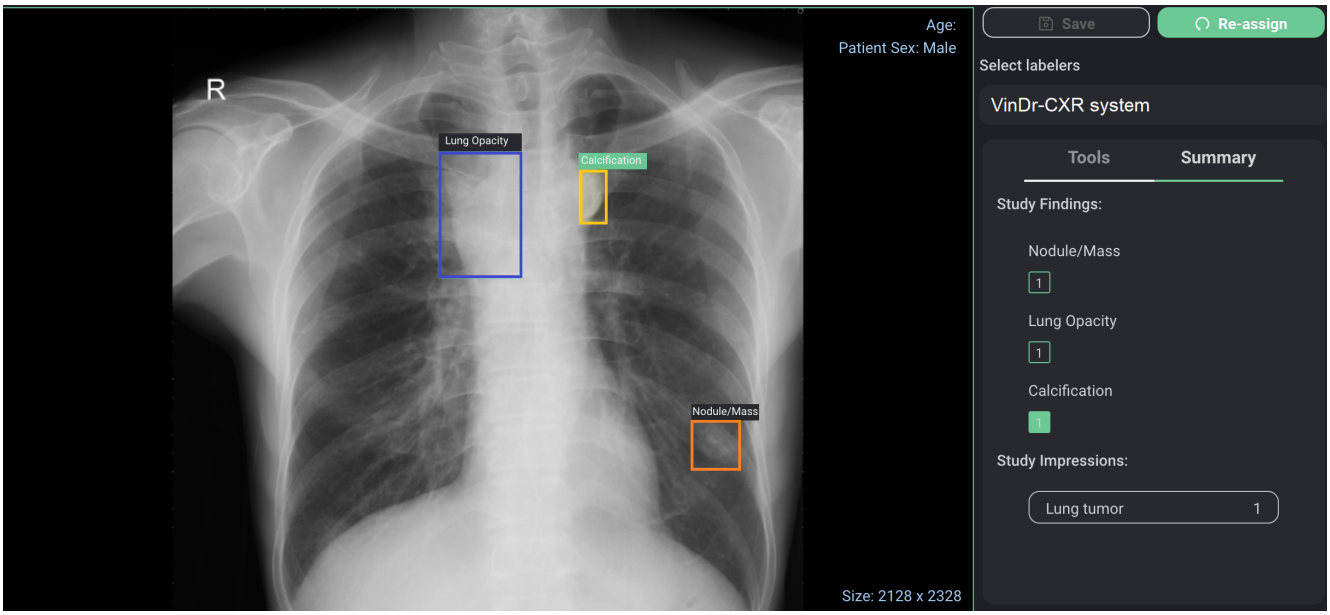

Figure 10. The clinical findings detected by the VinDr-CXR system are listed on the interface, and bounding box predictions are displayed on the image.

References

1. World Health Organization (WHO). Global tuberculosis report 2019. https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1. (accessed 05 March 2021).
2. Organization, W. H. Cancer statistics 2019. https://www.who.int/news-room/fact-sheets/detail/cancer. (accessed 04 March 2021).
3. Second edition forum of international respiratory societies. The global impact of respiratory disease. https://www.who.int/gard/publications/The_Global_Impact_of_Respiratory_Disease.pdf. (accessed 04 March 2021).
4. Corne, J. & Kumaran, M. *Chest X-ray made easy E-book* (Elsevier Health Sciences, 2015).
5. Delrue, L. *et al.* Difficulties in the interpretation of chest radiography. In *Comparative Interpretation of CT and Standard Radiography of the Chest*, 27–49 (Springer, 2011).
6. Fitzgerald, R. Error in radiology. *Clin. Radiol.* **56**, 938–946 (2001).
7. Manning, D. J., Ethell, S. & Donovan, T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. *The Br. J. Radiol.* **77**, 231–235 (2004).
8. Carmody, D. P., Nodine, C. F. & Kundel, H. L. An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception* **9**, 339–344 (1980).
9. Donald, J. J. & Barnard, S. A. Common patterns in 558 diagnostic radiology errors. *J. Med. Imaging Radiat. Oncol.* **56**, 173–178 (2012).
10. Gavelli, G. & Giampalma, E. Sensitivity and specificity of chest X-ray screening for lung cancer. *Cancer* **89**, 2453–2456 (2000).
11. Peng, J.-M. *et al.* Does training improve diagnostic accuracy and inter-rater agreement in applying the Berlin radiographic definition of acute respiratory distress syndrome? A multicenter prospective study. *Critical Care* **21**, 1–8 (2017).
12. Moncada, D. C. *et al.* Reading and interpretation of chest X-ray in adults with community-acquired pneumonia. *The Braz. J. Infect. Dis.* **15**, 540–546 (2011).
13. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60–88 (2017).
14. Rajpurkar, P. *et al.* ChexNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

15. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine* **15**, e1002686 (2018).

16. 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).

17. 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 (2020).

18. 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).

19. Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Medicine* **3**, 1–8 (2020).

20. Ting, D. S. W. *et al.* Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).

21. Zago, G. T., Andreão, R. V., Dorizzi, B. & Salles, E. O. T. Retinal image quality assessment using deep learning. *Comput. Biol. Medicine* **103**, 64–70 (2018).

22. Liu, Y. *et al.* A deep learning system for differential diagnosis of skin diseases. *Nat. Medicine* **26**, 900–908 (2020).

23. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).

24. 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 (2017).

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

26. Oakden-Rayner, L. Exploring the ChestXray14 dataset: problems. https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/ (2017). [Online; accessed 04-May-2021].

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

28. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *The BMJ* **368** (2020).

29. Amann, J., Blasimme, A., Vayena, E., Frey, D. & Madai, V. I. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med. Informatics Decis. Mak.* **20**, 1–9 (2020).

30. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **9**, e1312 (2019).

31. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 6105–6114 (PMLR, 2019).

32. Tan, M., Pang, R. & Le, Q. V. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 10781–10790 (2020).

33. Phillips, N. A. *et al.* CheXphoto: 10,000+ smartphone photos and synthetic photographic transformations of chest X-rays for benchmarking deep learning robustness. *arXiv preprint arXiv:2007.06199* (2020).

34. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).

35. Bunch, P., Hamilton, J., Sanderson, G. & Simmons, A. A free-response approach to the measurement and characterization of radiographic-observer performance. *J. Appl. Photogr. Eng.* **4**, 166–171 (1978).

36. FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. (accessed 5 May 2021).

37. Xie, Y., Chen, M., Kao, D., Gao, G. & Chen, X. CheXplain: Enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13 (2020).

38. Mauricio Reyes, S. P. C. A. S. F.-M. D. H. v. T.-K. R. M. S. R. W., Raphael Meier. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. *Radiology* **2** (2020).

39. Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M. & Hooman, A. An overview of principal component analysis. *J. Signal Inf. Process.* **4**, 173 (2013).

40. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine* **15** (2018).

41. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* **17**, 195 (2019).

42. Cohen, J. P., Hashir, M., Brooks, R. & Bertrand, H. On the limits of cross-domain generalization in automated X-ray prediction. *arXiv preprint arXiv:2002.02497* (2020).

43. 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).

44. European Parliament and Council of European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). https://gdpr-info.eu/ (2016). (Online; accessed 11 April 2021).

45. 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 May 2021).

46. Nguyen, H. Q. *et al.* VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *arXiv preprint arXiv:2012.15029* (2020).

47. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988 (2017).

48. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).

49. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378 (1971).

50. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* 159–174 (1977).

51. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

52. Wu, N. *et al.* Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Med. Imaging* (2019).

53. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).

54. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

55. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 740–755 (Springer, 2014).