Evaluating the Impact of an Explainable Machine Learning System on Interobserver Agreement in Chest Radiograph Interpretation

Hieu H. Pham
Coordinated Science Laboratory, UIUC
VinUni-Illinois Smart Health Center & CECS, VinUniversity
hieu.ph@vinuni.edu.vn

Ha Q. Nguyen
VinBigData JSC
v.HaNQ3@vinbigdata.org

Hieu T. Nguyen
Northeastern University
nguyen.trungh@northeastern.edu

Linh T. Le
Hanoi Medical University
linhdhyhn2017@gmail.com

Khanh Lam
108 Hospital
lamkhanh.himed@gmail.com

April 1, 2023

1 Introduction

The actual impact of AI systems on the diagnostic performance of radiologists in clinical practice remains unclear [12, 1, 11]. We developed an explainable deep learning system called VinDr-CXR that can classify a chest X-ray (CXR) into multiple thoracic diseases and localize critical findings on the image [8, 6, 15, 4, 9, 3, 2, 7, 5]. A prospective study was conducted to measure the clinical impact of the VinDr-CXR in assisting six experienced radiologists. The results indicated that when VinDr-CXR was used as a diagnosis-supporting tool, significantly improved the agreement between radiologists themselves with an increase of 1.5% in mean Fleiss' Kappa [10]. 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. This work has been accepted for publication in IEEE Access and its full-length version can be found in [9]. This is our short version submitted to the Midwest Machine Learning Symposium (MMLS 2023), Chicago, IL, USA (https://www.midwest-ml.org/2023/).

2 Our Approach

The proposed framework includes two major components. First, an image-level classification network [13] accepts a CXR scan as input and predicts whether it could be normal or abnormal. Second, a lesion-level detection network [13, 14] receives an abnormal CXR scan as input from the classifier and provides the location of abnormal findings via bounding box predictions. The core of the VinDr-CXR system is based on state-of-the-art DL networks for image classification and object detection tasks. VinDr-CXR was trained on 51,485 CXR scans with radiologist-provided bounding box annotations [4]. The actual impact of the VinDr-CXR was evaluated through a reader study (*N* = 400). The inter-rater agreement among radiologists as well as the rate of agreement between VinDr-CXR and radiologists are then assessed the Cohen's Kappa metric. On a validation set of 3,000 CXR studies. The system reported a mean AUROC of 0.967 (95% CI: 0.958, 0.975) for the classification task. For the detection, it achieved a sensitivity of 80.2% (81.4, 84.9) at 1.0 false-positive marks per image. The FROC of the VinDr-CXR system was 78.36% (76.46, 80.16).

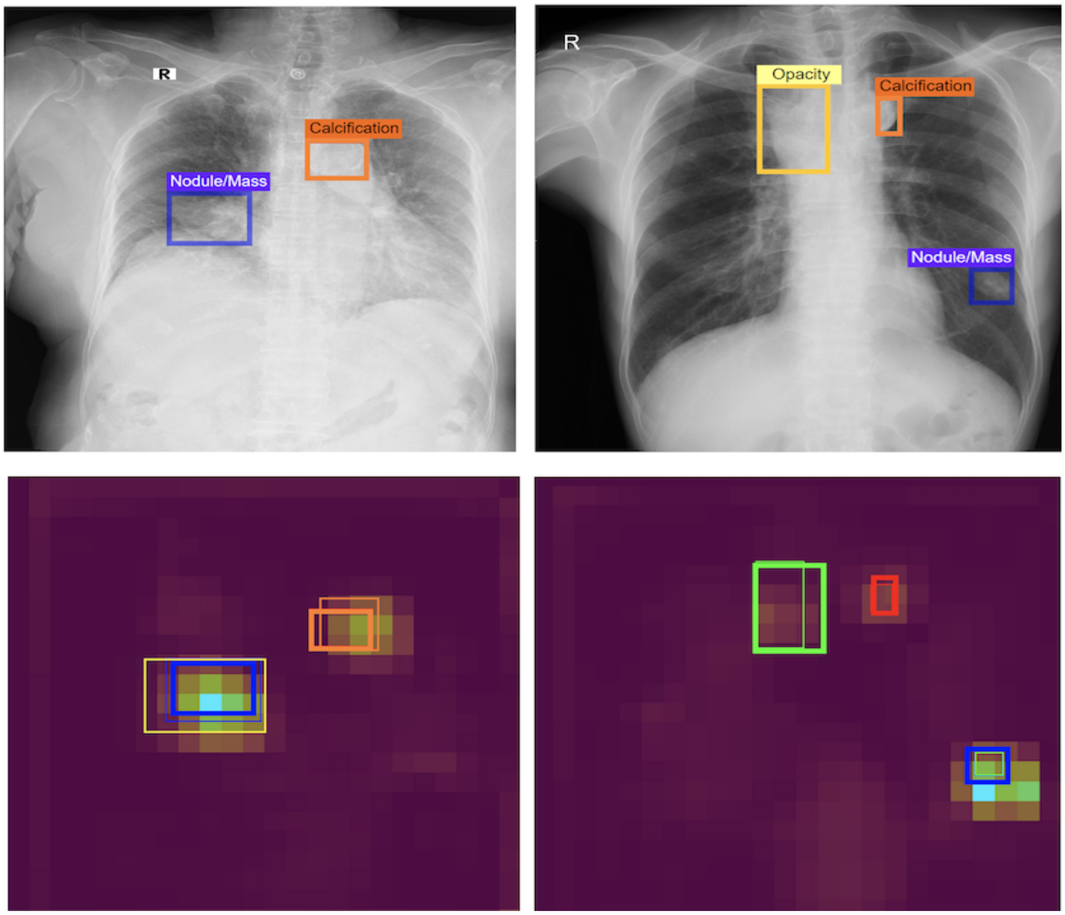

Figure 1: VinDr-CXR localizes critical findings on the image.

2.1 Impact of VinDr-CXR in clinical practice

We recruited a group of six board-certified radiologists from 108 hospital (H108) and Hanoi Medical University (HMUH) to participate in our observer performance test. 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. They considered the model's prediction and modified the diagnostics. Our experiments showed that in the second read, with the support of the VinDr-CXR system, agreement among three H108's radiologists was moderate with a Fleiss' Kappa of 0.545 (0.465, 0.625), corresponding to a 3.0% improvement in Fleiss' Kappa compared to the first read. Additionally, we found that the rate of VinDr-CXR agreement with the participating radiologists was slightly higher than the rate of agreement among radiologists. The agreement between each radiologist and the system was remarkably increased by 3.3% in mean Cohen's Kappa.

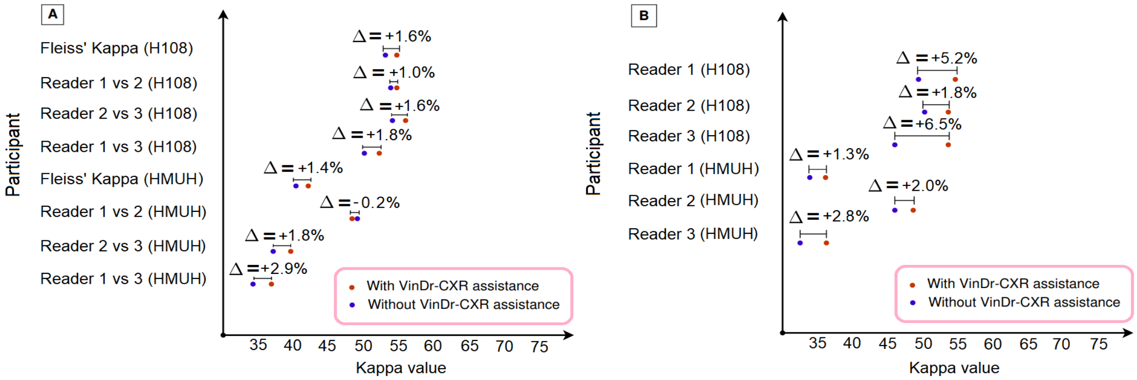

Figure 2: Change in inter-radiologist agreement before and after consulting the VinDr-CXR predictions.

3 Conclusion

This study showed that an accurate and explainable deep learning system is able to improve interobserver agreement in the interpretation of chest radiograph. Further research is needed to validate the model prospectively and determine its utility in clinical settings.

References

[1] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[2] K. H. Le, T. V. Tran, H. H. Pham, H. T. Nguyen, T. T. Le, and H. Q. Nguyen. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *IEEE Access*, 11:14105–14114, 2023.

[3] H. N. Nguyen, H. Pham, T. Tran, T. Nguyen, and Q. H. Nguyen. Vindr-pcxr: An open, large-scale chest radiograph dataset for interpretation of thoracic diseases in children. *medRxiv*, pages 2022–03, 2022.

[4] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. Tong, D. H. Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

[5] N. H. Nguyen, H. Q. Nguyen, N. T. Nguyen, T. V. Nguyen, H. H. Pham, and T. N.-M. Nguyen. A clinical validation of vindr-cxr, an ai system for detecting abnormal chest radiographs. *arXiv preprint arXiv:2104.02256*, 2021.

[6] N. H. Nguyen, H. Q. Nguyen, N. T. Nguyen, T. V. Nguyen, H. H. Pham, and T. N.-M. Nguyen. Deployment and validation of an AI system for detecting abnormal chest radiographs in clinical settings. *Frontiers in Digital Health*, page 130, 2022.

[7] T. Nguyen, T. M. Vo, T. V. Nguyen, H. H. Pham, and H. Q. Nguyen. Learning to diagnose common thorax diseases on chest radiographs from radiology reports in vietnamese. *Plos one*, 17(10):e0276545, 2022.

[8] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021.

[9] H. H. Pham, H. Q. Nguyen, H. T. Nguyen, L. T. Le, and L. Khanh. An accurate and explainable deep learning system improves interobserver agreement in the interpretation of chest radiograph. *IEEE Access*, 10:104512–104531, 2022.

[10] G. Rücker, T. Schimek-Jasch, and U. Nestle. Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using fleiss' kappa. *Methods of information in medicine*, 51(06):489–494, 2012.

[11] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

[12] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

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

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

[15] T. T. Tran, H. H. Pham, T. V. Nguyen, T. T. Le, H. T. Nguyen, and H. Q. Nguyen. Learning to automatically diagnose multiple diseases in pediatric chest radiographs using deep convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3314–3323, 2021.