Radiology: Artificial Intelligence

Collaborative Integration of AI and Human Expertise to Improve Detection of Chest Radiograph Abnormalities

Akash Awasthi[1] (http://orcid.org/0000-0002-8383-0637)

Ngan Le, PhD[2] (http://orcid.org/0000-0003-2571-0511)

Zhigang Deng, PhD[3] (http://orcid.org/0000-0002-0452-8676)

Carol C. Wu, MD[4] (http://orcid.org/0000-0003-1005-0995)

Hien Van Nguyen, PhD[5]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

https://doi.org/10.1148/ryai.240277

Purpose: To develop a collaborative AI system that integrates eye gaze data and radiology reports to improve diagnostic accuracy in chest radiograph interpretation by identifying and correcting perceptual errors.

Materials and Methods: This retrospective study utilized public datasets REFLACX and EGD-CXR to develop a collaborative AI solution, named Collaborative Radiology Expert (CoRaX). It employs a large multimodal model to analyze image embeddings, eye gaze data, and radiology reports, aiming to rectify perceptual errors in chest radiology. The proposed system was evaluated using two simulated error datasets featuring random and uncertain alterations of five abnormalities. Evaluation focused on the system's referral-making process, the quality of referrals, and its performance within collaborative diagnostic settings.

Results: In the random masking-based error dataset, 28.0% (93/332) of abnormalities were altered. The system successfully corrected 21.3% (71/332) of these errors, with 6.6% (22/332) remaining unresolved. The accuracy of the system in identifying the correct regions of interest for missed abnormalities was 63.0% [95% CI: 59.0%, 68.0%], and 85.7% (240/280) of interactions with radiologists were deemed satisfactory, meaning that the system provided diagnostic aid to radiologists. In the uncertainty-masking-based error dataset, 43.9% (146/332) of abnormalities were altered. The system corrected 34.6% (115/332) of these errors, with 9.3% (31/332) unresolved. The accuracy of predicted regions of missed abnormalities for this dataset was 58.0% [95% CI: 55.0%, 62.0%], and 78.4% (233/297) of interactions were satisfactory.

Conclusion: The CoRaX system can collaborate efficiently with radiologists and address perceptual errors across various abnormalities in chest radiographs.

The proposed system, CoRaX, demonstrated potential to aid radiologists in detection of abnormalities on chest radiographs by identifying and correcting perceptual errors.

Abbreviations:

CoRaX = Collaborative Radiology Expert, LMM = Large multimodal model, TR = True Referral, FR = False Referral, FD = False Deferral, TD = True Deferral, PECR = Perceptual Error Correction Rate, ODER = Over-Diagnosis Error Rate, STARE = Spatio-Temporal Abnormal Region Extractor, MAF = Missing Abnormality Finder

Key Points:

- A collaborative AI system, CoRaX, was developed to aid radiologists in the detection of abnormalities on chest radiographs.
- CoRaX corrected 21.3% (71 of 332) and 34.6% (115/332) of errors, defined as alterations in abnormalities, in the two simulated error datasets.
- In a collaborative diagnostic setting, the system had a non-zero interaction score in 85.7% (240/280) and 78.4% (233/297) of interactions for the error datasets, indicating its potential to aid radiologists in diagnostic decision-making.

Artificial intelligence (AI) systems play an increasingly important role in health care decision-making, particularly in diagnostic processes (1–4). However, their integration faces challenges, especially in standalone systems that generally overlook human interaction (5,6). Use of such systems may lead to decreased diagnostic accuracy due to varying levels of trust among medical professionals (7–10). Collaborative Intelligence offers a promising solution by emphasizing human-AI synergy to enhance accuracy, notably in radiology.

A key challenge in radiology is the prevalence of perceptual errors, which significantly impact diagnostic accuracy. These errors occur when radiologists fail to detect or correctly interpret abnormalities due to visual oversights during the initial interpretation (11,12). Research indicates that perceptual errors are responsible for most diagnostic mistakes, ranging from 60 to 80% (13). Addressing these perceptual errors is critical to improve diagnostic performance (14). While there are various insights into the causes and potential manual solutions for perceptual errors (14), no AI-based solutions currently address these errors based on the individual visual search patterns of radiologists.

Eye gaze serves as a valuable sensing modality in human-computer interaction (15), fostering dynamic collaboration between radiologists and AI systems (16). Existing studies have shown a direct correlation between eye movements and radiologist's diagnostic decisions (17–19), making eye-tracking data crucial to understand the visual search process (20–26) and reducing diagnostic errors (18,19). Furthermore, eye gaze recording is non-intrusive and can be seamlessly integrated into clinical workflows.

This study aims to develop a personalized AI system, named Collaborative Radiology Expert (CoRaX). It integrates eye gaze data and radiology reports to improve diagnostic accuracy of radiologists in chest radiograph (CXR) interpretation by identifying and correcting perceptual errors. Incorporation of eye gaze data, which reflects the unique patterns of how a radiologist

views and analyzes CXRs, into the AI model allows for the development of a personalized system that provides highly individualized referrals.

Materials and Methods

The retrospective study was approved by the Institutional Review Board (approval number for this study is STUDY00003659). conducted in accordance with the ethical standards of the responsible committee on human experimentation and with the Helsinki Declaration of 1975, as revised in 2000.

Datasets

This study utilized the public datasets EGD-CXR (27) and REFLACX (28). The EGD-CXR dataset comprises 1,071 CXRs reviewed by radiologists using an eye-tracking system. The REFLACX dataset encompasses 2440 cases with synchronized eye-tracking and speech transcription pairs, annotated by radiologists. We generated fixation heatmaps overlaid on the CXRs by leveraging eye gaze data to provide a dynamic representation of the gaze movements over the CXR images. The two datasets belong to two different radiologists, and we combined them and created a training set and a test set to train the STARE module. The speech transcription data obtained from EGD-CXR and REFLACX contain detailed radiology reports with word alignments for CXR images. By merging the transcriptions from REFLCAX and EGD-CXR, we generated a final JSON file after preprocessing that includes comprehensive reports and associated timesteps. This compilation is crucial for training the STARE module.

System Overview

This study specifically addresses visual misses caused by recognition and decision issues, categorizing them as perceptual errors when radiologists fail to mention or recognize abnormalities (29). One conventional approach to reducing perceptual errors is the Double Reading system, where a second reader reviews the study (14). Our system functions as a virtual second reader or postinterpretation tool to mitigate perceptual errors. CoRaX operates as a postinterpretation system, where radiologists submit radiographic images, reports, and eye gaze data (Fig 1). CoRaX then generates referrals for further assessment by radiologists, forming a collaborative framework between them and the system, with eye gaze data aiding in understanding radiologists' cognitive processes. CoRaX comprises two crucial modules: 1) Missed Abnormality Finder (MAF) and 2) Spatio-Temporal Abnormal Region Extractor (STARE) (Fig 2). Additionally, it involves specific set operations to identify perceptual errors. Its primary focus is not to generate full radiology reports but rather to identify key findings or abnormalities mentioned in the Chexpert dataset (30) overlooked by the radiologist.

The MAF module is tasked with summarizing the radiology report and identifying any missing abnormalities. If the radiologist fails to diagnose an abnormality in the CXR image, this module appends the missing abnormality to the summarized radiology report. MAF is created by combining the functionalities of the Chexpert Labeler (30) and ChexFormer.

Chexpert-Labeler (30) is a rule-based natural language processing (NLP) tool used to automatically annotate radiology reports for CXRs with a set of predefined labels. Developed as part of the Chexpert dataset (30), it identifies the presence, absence, or uncertainty of specific medical conditions (such as pneumonia, atelectasis, and cardiomegaly) by parsing and analyzing the textual content of the reports. We have used the Chexpert-Labeler (30) to summarize the radiology report into the Chexpert labels.

ChexFormer serves as the fundamental engine of our proposed system—a multilabel transformer classifier designed to predict multiple labels corresponding to a given chest X-ray (CXR) image (31). Pretrained on the Chexpert dataset (30), ChexFormer demonstrates proficiency in multilabel classification tasks. Refer to the detailed architectural description of the ChexFormer and MAF module in the supplementary section.

The STARE module is a pivotal component of our system. It predicts the temporal grounding (ie, timestamp) for each abnormality in the corrected and summarized radiology report. Inspired by a dense video captioning task in computer vision (32,33). STARE takes the MAF module's output (ie, a corrected and summarized radiology report) and the eye gaze video to predict timestamps for each abnormality in the report, as depicted in Figure 2. The detailed overview of the STARE module is presented in the supplementary file.

The final referral generation process addresses visual misses by comparing label sets. Set A represents the original report, while Set B corresponds to the corrected prediction. A set difference operation is applied to identify discrepancies between the two sets (Fig 2). A detailed explanation of the set difference and referral generation logic can be found in the supplementary material. We provide the link to the source code and dataset in this link: https://github.com/a04101999/CoRaX-Collaborative-radiology-xpert-

Training and Testing

To manage computational costs while leveraging public datasets, we trained each module independently. ChexFormer was trained on approximately 4,000 chest X-rays and tested on 1,000 images from the CheXpert dataset (30). During training, it received both label embeddings and image features; during inference, only image features were used to predict labels. CheXpert includes 14 labels with values 0 (absent), 1 (present), and-1 (uncertain). We treated uncertain (−1) labels as positive (1), aligning with our goal to flag potentially missed abnormalities.

The STARE module is trained on a merged dataset comprising the REFLACX and EGD-CXR datasets, resulting in a total of 3511 samples. Among these samples, 2,969 are designated for training, 271 for validation, and 271 for testing. Inputs included fixation heatmap videos and temporally grounded abnormality sequences derived from summarized reports. The model was trained for 300 epochs using the Adam optimizer with a batch size of 2 on 8 Tesla GPUs. Detailed standalone performance of the STARE and ChexFormer modules is included in the supplementary section.

Error Dataset

The principal aim of our system was to identify and rectify perceptual errors. Since real-time datasets with perceptual errors are unavailable, we created error datasets by simulating errors from the test set. CoRaX was evaluated on these datasets to assess its effectiveness in identifying and rectifying perceptual errors. The summarized radiology report for images in the test set (n = 271) was altered deliberately to introduce diagnostic errors across different abnormalities in different regions of radiographs. Our system is specifically designed to correct errors for five abnormalities (Table 1): cardiomegaly, pleural effusion, atelectasis, lung opacity, and edema. Importantly, all errors evaluated and created in this study are limited to these five abnormalities. Error dataset is created in two ways: 1) Based on the random masking of the abnormalities, 2) Masking based on the Uncertain abnormalities

Random Masking of Abnormalities

Approximately 28.0% (93/332) of cases in the test set were randomly adjusted or masked to simulate errors. This approach involved making alterations without specific criteria to ensure that the system could assist both inexperienced radiologists, such as trainees, and experienced professionals. We assume that random masking is a valuable approach for simulating perceptual errors due to its broad coverage and generalization. By introducing errors in an unsystematic fashion, this technique allows for a comprehensive evaluation of the system's performance across a diverse array of abnormalities and scenarios. However, it may not accurately represent real-world perceptual errors. To address this limitation, we developed an additional dataset as described below.

Masking Based on Uncertain Abnormalities

Another error dataset was created by altering abnormalities based on uncertainty labels (−1). This approach was applied to the same test set used for random masking. For abnormalities with uncertain labels (eg, Cardiomegaly, Pleural Effusion, Atelectasis, Lung Opacity, Edema), we introduced simulated errors to better represent real-world scenarios where radiologists might overlook or misinterpret abnormalities due to complexity. This method aims to reflect the challenges faced in clinical practice. Table 1 shows the percentage of abnormalities introduced into the dataset. Approximately 44.0% (146/332) of the cases were adjusted to simulate errors based on uncertain labels. For instance, Cardiomegaly was missed in only 6.1% (4/65) of cases in the uncertain masking scenario, illustrating that it is generally easier to diagnose and less prone to confusion.

When we refer to "altered" cases, we mean that specific abnormalities were either masked or negated, representing perceptual errors from recognition or decision-making. The comparison between random masking and uncertain masking is crucial for understanding how different types of simulated errors impact system performance. Random masking provides a baseline to evaluate overall error correction capability, while uncertain masking offers insights into how the system handles errors linked to diagnostic uncertainty. As detailed in Table 1, various abnormalities in the test data were modified at different percentages to simulate these errors.

Referral Evaluation Metrics

We evaluated CoRaX's referral accuracy using metrics that quantify its ability to identify and correct perceptual errors while minimizing false alarms. These include True Referral (TR), False Referral (FR), False Deferral (FD), and True Deferral (TD), which together characterize referral acceptance and rejection patterns. From these, we derive the Perceptual Error Correction Rate (PECR) and the Overdiagnosis Error Rate (ODER) to assess correction sensitivity and over-referral tendency, respectively. To evaluate the spatial precision of referrals, we compute Interpretable Referral Accuracy (IRA), which combines referral acceptance with the overlap of predicted regions and ground truth using the Intersection over Union (IoU) (34). Finally, we introduce an Interaction Score to assess the quality of system–radiologist interactions, distinguishing between referral-based and deferral-based cases. An interaction is considered fully effective when both diagnostic and spatial correctness are achieved. A detailed description of these metrics is provided in the supplementary material.

Statistical Analysis

All analyses are performed using Python version 3.8. PECR and ODER metrics are calculated to evaluate the accuracy of Referrals and Deferrals. The 95% confidence intervals (CIs) for the PECR and ODER were also calculated using the bootstrap method with 1000 bootstrap samples. In cases where a metric achieved a perfect score (eg, 100%), the Wilson score interval was used instead, with a significance level of $\alpha = 0.05$, yielding 95% CIs. Interpretable Referral Accuracy (IRA) metric based on the Jaccard Index, also called IoU score, is used to assess the system's ability to predict the region of interest for each missed abnormality. To provide robust variability estimates, 95% confidence intervals (CIs) for the Interpretable Referral Accuracy were calculated using the bootstrap method with 1000 bootstrap samples. Separate 95% CIs were also calculated for each abnormality-specific true referral using the same bootstrap approach. Additionally, we have plotted the histograms to analyze the distribution of the Interaction Score, assessing both referral and non-referral interactions to provide a comprehensive view of the system's performance across various interactions. The 95% confidence intervals (CIs) for the Interaction Score were also calculated using the bootstrap method with 1000 bootstrap samples. Details on the calculation of confidence Intervals (CIs) using the bootstrap method are provided in the supplementary file. No statistical significance tests were conducted in this study.

Results

Illustration of Referrals

CoRaX demonstrated the ability to identify clinically relevant missed findings and generate appropriate referrals. It successfully detected multiple abnormalities overlooked in original radiology reports, including bilateral findings such as pleural effusion and pulmonary edema, which carry significant clinical implications if left unrecognized. Additionally, the system addressed uncertainty-related reporting errors by referring cases in which uncertain findings were omitted in erroneous reports (Fig 3).

Referral Evaluation

The Random Masking-based error dataset contains approximately 271 samples, with 93 abnormalities missed across various regions in the radiographs. It is important to note that the 93 missed abnormalities represent the number of abnormal conditions missed (TR + FD) rather than the number of individual cases, as a single case can contain multiple abnormalities. For this dataset, CoRaX achieved an overall PECR of 76.3% (71/93) [95% CI: 45.4%, 100.0%] and an overall ODER of 4.6% (9/195) [95% CI: 2.0%, 7.6%]. Table 2 shows the abnormality-specific PECR and ODER.

The Uncertainty Masking-based error dataset also contains around 271 samples, with 146 abnormalities missed. For this dataset, CoRaX achieved an overall PECR of 78.7% (115/146) [95% CI: 59.2%, 100.0%] and an overall ODER of 6.4% (10/155) [95% CI: 3.6%, 11.2%]. Abnormality-specific PECR and ODER are provided in Table 3.

Overall, CoRaX demonstrates a higher PECR and lower ODER for Cardiomegaly compared with other lung diseases in both datasets, indicating that Cardiomegaly is more easily detectable than other conditions. This is because cardiomegaly has a defined region near the heart, but other lung diseases can occur in any area in the lungs.

Evaluation of Referrals Based on the Region of Interest

The overall IRA for all true referrals showed a mean score of 63.0% [95% CI: 59.0%, 68.0%] in the Random Masking-Based Error Dataset (Fig 4(1A)) and 58.0% [95% CI: 55.0%, 62.0%] in the Uncertainty Masking-Based Error Dataset (Fig 4(2A)). For abnormality-specific performance, Cardiomegaly consistently demonstrated the highest IRA, with a mean of 83.0% [95% CI: 74.0%, 90.0%] in the Random Masking-Based Error Dataset and 84.0% [95% CI: 69.0%, 100.0%] in the Uncertainty Masking-Based Error Dataset (Figs 4(1B) and 4(2B)). Detailed mean IRA scores for other abnormalities are provided in the supplementary material. For overall performance, histogram plots of IRA scores show that 81.2% (65/80) of referrals in the Random Masking-Based Error Dataset had an IRA greater than 0, with 18.7% (15/80) scoring 0, while 73.6% (92/125) of referrals in the Uncertainty Masking-Based Error Dataset demonstrated non-zero scores, with 26.4% (33/125) scoring 0 (Figs 4(1C) and 4(2C)).

Evaluation of Overall Interaction and Diagnostic Accuracy

In the Random Masking-Based Error Dataset (Table 4), approximately 71.4% (200/280) of the interactions resulted in no referrals, with around 63.5% (178/280) of these decisions being accurate, and 7.8% (22/280) of missed correcting radiologist's perceptual errors. Regarding referral-based interactions, 28.5% (80/280) of interactions are referral based, with around 25.3% (71/280) were correct and 3.2% (9/280) are incorrect, leading to direct rejection by radiologists. The system achieved a non-zero Interaction score, indicating its ability to provide diagnostic aid to radiologists, in 85.7% (240/280) of interactions. The mean Interaction score for the Random masking-based error dataset is 0.78[95% CI: 0.74, 0.83] (Fig 5(A)).

In the Uncertainty Masking-Based Error Dataset (Table 5), approximately 58.2% (173/297) of interactions resulted in no referrals, with about 47.8% (142/297) being accurate and 10.4% (31/297) missing perceptual errors. For referral-based interactions, 42.0% (125/297) involved referrals, with 38.7% (115/297) correct and 3.3% (10/297) incorrect, leading to rejections by radiologists. The system achieved a non-zero Interaction score, indicating its ability to provide diagnostic aid to radiologists, in 78.4% (233/297) of interactions. The mean Interaction score for the Uncertainty masking–based error dataset is 0.66[95% CI: 0.61, 0.70] (Fig 5(B)).

Discussion

This study introduces CoRaX, a collaborative AI system designed to address perceptual errors in chest radiograph interpretation by integrating gaze data and radiologist interaction patterns. CoRaX was evaluated using two simulated error datasets—random masking-based and uncertainty masking-based—to reflect common perceptual oversights. The system corrected 21.3% (71/332) of errors in the random masking dataset and 34.6% (115/332) in the uncertainty masking dataset, with particularly strong performance in identifying missed Cardiomegaly cases. It achieved mean IRA scores of 63.0% [95% CI: 59.0, 68.0%] and 58.0% [95% CI: 55.0, 62.0%, and provided diagnostic aid in 85.7% (240/280) and 78.4% (233/297) of interactions, respectively, as measured by the Interaction Score.

Prior research in radiology AI has largely centered on fully automated systems for abnormality detection (30) or report generation (36), rather than targeting perceptual error correction. While these systems offer valuable diagnostic support, their autonomous nature may contribute to over-reliance or skepticism among clinicians, particularly when lacking interpretability (9). CoRaX differs fundamentally by functioning as a collaborative assistant: it supports radiologists by identifying potential perceptual oversights and offering interpretable, gaze-informed feedback. Metrics such as Interpretable Referral Accuracy quantify the system's ability to redirect attention effectively, while the Interaction Score captures its broader diagnostic utility. This framework emphasizes human-AI synergy over substitution, fostering more actionable and trustworthy clinical support.

This study has several limitations. First, perceptual errors were synthetically introduced into the datasets. Although these were designed to resemble real-world oversights through manual curation of erroneous samples, they do not fully capture the diversity of perceptual errors present in clinical radiology (29,35). Second, minor misalignments between gaze data and transcription timestamps led to slight inaccuracies in the regions predicted by the STARE module. Furthermore, the current study focuses primarily on technical system development; direct real-world validation with radiologists remains a future goal. Nonetheless, CoRaX shows promise for educational use, particularly in training less experienced radiologists. Its modular design facilitates ongoing research and adaptation, supporting future refinement and personalization.

In conclusion, CoRaX is a novel AI system that addresses perceptual errors in radiology through multimodal analysis of gaze and interaction data. By developing and evaluating two simulated error datasets, we demonstrated its capacity to correct missed findings and enhance diagnostic performance. While the current work emphasizes system development, the modular

architecture enables future enhancements, such as replacing the multilabel classifier (eg, ChexFormer) with more advanced models. This approach lays the groundwork for robust, error-resistant AI systems and paves the way for future clinical trials and broader adoption.

Author affiliations:

[1] Department of Electrical and Computer Engineering, University of Houston, 4222 Martin Luther King Blvd, Cullen College of Engineering Building-1, Rm N368, Houston, TX 77204

[2] Department of Computer Science & Computer Engineering, University of Arkansas, Fayetteville, Ark

[3] Department of Computer Science, University of Houston, Houston, Tex

[4] Department of Thoracic Imaging, Division of Diagnostic Imaging, The University of Texas MD Anderson Cancer Center, Houston, Tex

[5] Department of Electrical and Computer Engineering, University of Houston, Houston, Tex

Received XXX; revision requested XXX; revision received XXX; accepted XXX

Address correspondence to: J.Z. (e-mail: akashcseklu123@gmail.com).

Funding: Work supported by NIH 1R01CA277739

Author contributions: Guarantor of integrity of entire study, **A.A., N.L., H.V.N.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **A.A., N.L., H.V.N.**; clinical studies, **A.A., N.L., H.V.N.**; experimental studies, all authors; statistical analysis, **A.A., N.L., H.V.N.**; and manuscript editing, **A.A., N.L., Z.D., H.V.N.**

Disclosures of conflicts of interest: **A.A.** National Institutes of Health under grant 1R01CA277739. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. **N.L.** National Institutes of Health (NIH) 1R01CA277739-01. **Z.D.** No relevant relationships. **C.C.W.** NIH/University of Houston grant (co-investigator). **H.V.N.** National Cancer Institute (5R01CA277739) (payment made to institution.

References

1. Topol E. Deep medicine: how artificial intelligence can make healthcare human again. London, United Kingdom: Hachette UK; 2019.

2. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380(14):1347–1358.

3. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J 2019;6(2):94–98.

4. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375(13):1216–1219.

5. Gilbert F. Balancing human and AI roles in clinical imaging. Nat Med 2023;29(7):1609–1610.

6. Dvijotham KD, Winkens J, Barsbey M, et al. Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians. Nat Med 2023;29(7):1814–1820.

7. Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR & RSNA. Can Assoc Radiol J 2024;75(2):226–244.

8. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020;22(6):e15154.

9. Kundu S. Measuring trustworthiness is crucial for medical AI tools. Nat Hum Behav 2023;7(11):1812–1813.

10. Quinn TP, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. J Am Med Inform Assoc 2021;28(4):890–894.

11. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. RadioGraphics 2015;35(6):1668–1676.

12. Garland LH. On the scientific evaluation of diagnostic procedures: Presidential address thirty-fourth annual meeting of the Radiological Society of North America. Radiology 1949;52(3):309–328.

13] Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D. Interpretive error in radiology. AJR Am J Roentgenol 2017;208(4):739–749.

14. Degnan AJ, Ghobadi EH, Hardy P, et al. Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. Acad Radiol 2019;26(6):833–845.

15. Chen H, Zendehdel N, Leu MC, Yin Z. Real-time human-computer interaction using eye gazes. Manuf Lett 2023;35:883–894.

16. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic component of image perception in mammogram interpretation: gaze-tracking study. Radiology 2007;242(2):396–402.

17. Tourassi G, Voisin S, Paquit V, Krupinski E. Investigating the link between radiologists' gaze, diagnostic decision, and image content. J Am Med Inform Assoc 2013;20(6):1067–1075.

18. Samuel S, Kundel HL, Nodine CF, Toto LC. Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. Radiology 1995;194(3):895–902.

19. Manning DJ, Ethell SC, Donovan T. Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph. Br J Radiol 2004;77(915):231–235.

20. Krupinski EA. Influence of experience on scanning strategies in mammography. In Medical Imaging 1996: Image Perception 1996 Mar 27 (Vol 2712, pp. 95–101). SPIE. https://doi.org/10.1117/12.236845.

21. Mugglestone MD, Gale AG, Cowley HC, Wilson AR. Defining the perceptual processes involved with mammographic diagnostic errors. In Medical Imaging 1996: Image Perception 1996 (Vol 2712, pp. 71–77). SPIE. https://doi.org/10.1117/12.236862.

22. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol 1996;3(2):137–144.

23. Nodine CF, Mello-Thoms C, Weinstein SP, et al. Blinded review of retrospectively visible unreported breast cancers: an eye-position analysis. Radiology 2001;221(1):122–129.

24. Mello-Thoms C, Hardesty L, Sumkin J, et al. Effects of lesion conspicuity on visual search in mammogram reading1. Acad Radiol 2005;12(7):830–840.

25. Mello-Thoms C. How does the perception of a lesion influence visual search strategy in mammogram reading? Acad Radiol 2006;13(3):275–288.

26. Mello-Thoms C, Britton C, Abrams G, et al. Head-mounted versus Remote Eye Tracking of Radiologists Searching for Breast Cancer: A Comparison1. Acad Radiol 2006;13(2):203–209.

27. Karargyris A, Kashyap S, Lourentzou I, et al. Eye gaze data for chest x-rays. PhysioNet. 2020. https://doi.org/10.13026/QFDZ-ZR67.

28. Bigolin Lanfredi R, Zhang M, Auffermann WF, et al. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Sci Data 2022;9(1):350.

29. Gefter WB, Post BA, Hatabu H. Commonly missed findings on chest radiographs: causes and consequences. Chest 2023;163(3):650–661.

30. Irvin J, Rajpurkar P, Ko M, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. 2019;33(01):590–597. In Proceedings of the AAAI conference on artificial intelligence. https://doi.org/10.1609/aaai.v33i01.3301590.

31. Lanchantin J, Wang T, Ordonez V, Qi Y. General multi-label image classification with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 16478–16488). https://doi.org/10.1109/CVPR46437.2021.01621.

32. Yang A, Nagrani A, Seo PH, et al. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition 2023 (pp. 10714–10726). https://doi.org/10.1109/CVPR52729.2023.01032.

33. Iashin V, Rahtu E. Multi-modal dense video captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops 2020 (pp. 958–959). https://doi.org/10.1109/CVPRW50498.2020.00487.

34. Rahman MA, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In International symposium on visual computing 2016 Dec 10 (pp. 234–244). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50835-1_22.

35. Vosshenrich J, Nesic I, Cyriac J, Boll DT, Merkle EM, Heye T. Revealing the most common reporting errors through data mining of the report proofreading process. Eur Radiol 2021;31(4):2115–2125.

36. Zhang Y, Wang X, Xu Z, Yu Q, Yuille A, Xu D. When radiology report generation meets knowledge graph. 2020;34(07):12910–12917. In Proceedings of the AAAI conference on artificial intelligence. https://doi.org/10.1609/aaai.v34i07.6989.

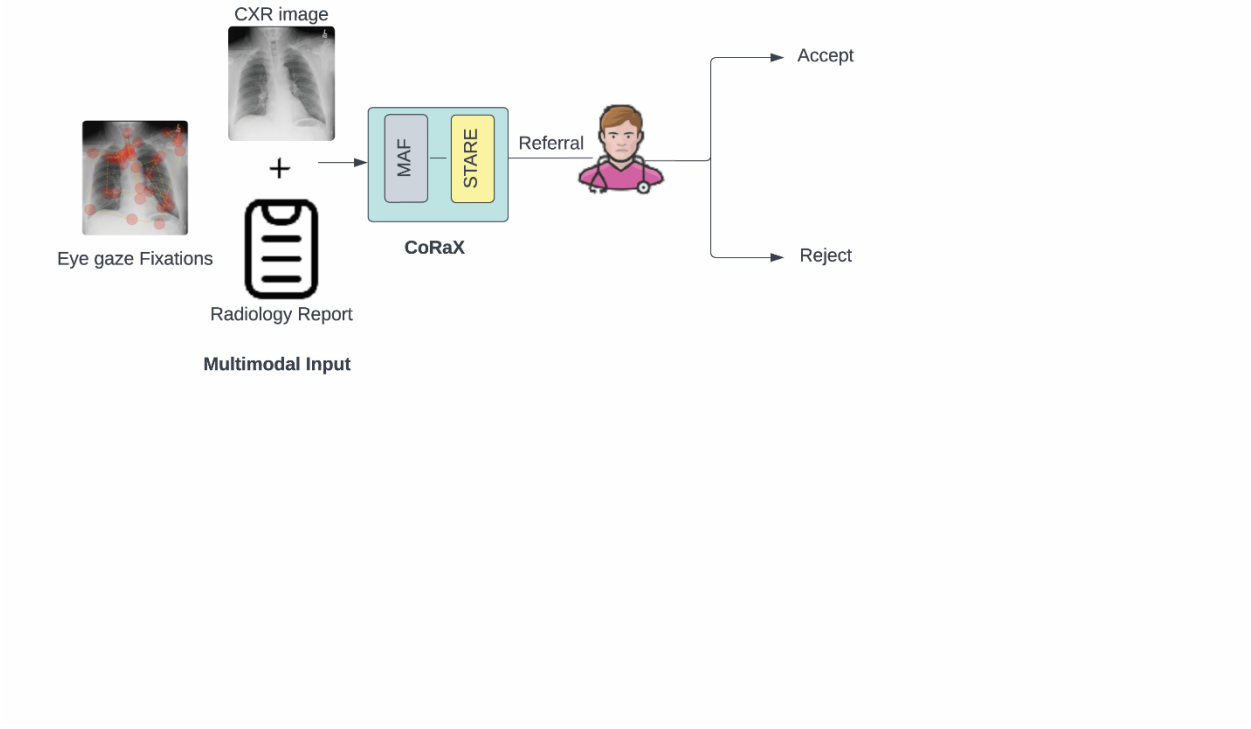


Figure 1: An overview of our innovative collaborative system, CoRaX. Our system seamlessly integrates radiology reports, eye gaze data, and chest radiographs (CXR) to offer targeted

recommendations. Then the radiologist uses these recommendations and either accepts them or rejects them. CXR = Chest X-ray.

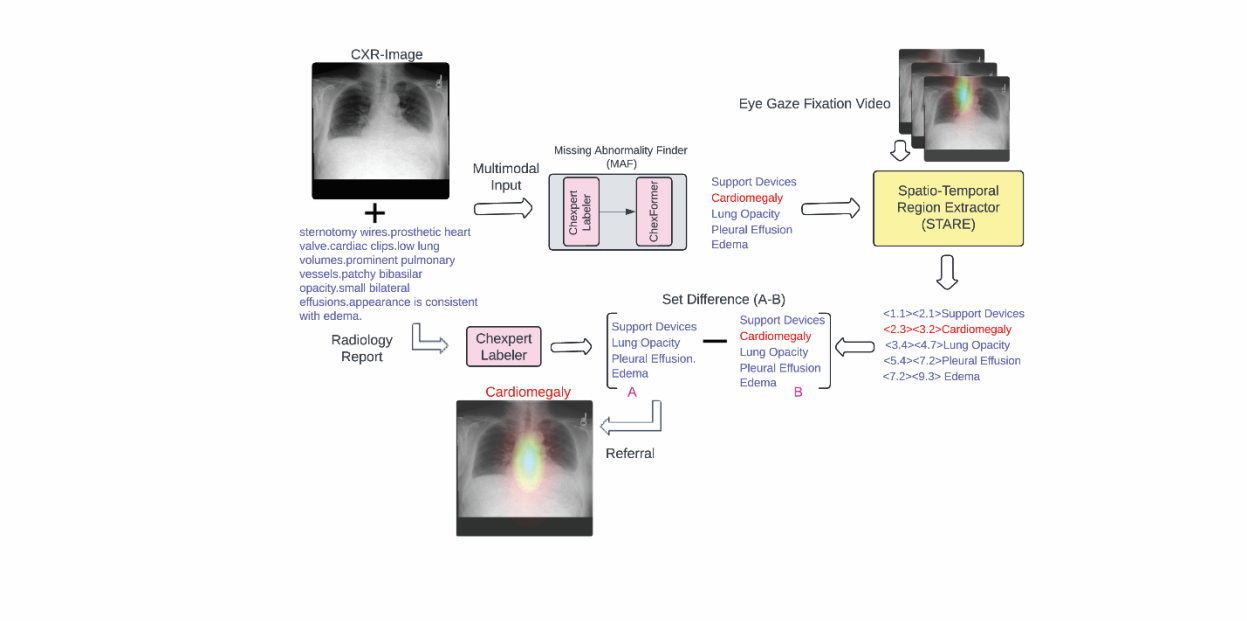
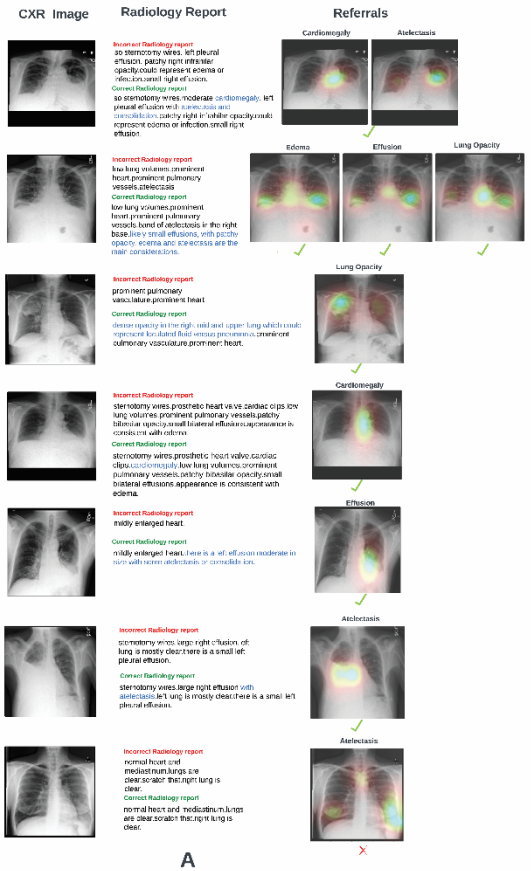


Figure 2: Overview of CoRaX architecture, consisting of two main modules. The MAF module is dedicated to identifying abnormalities that may have been missed, while the STARE module focuses on precisely locating the corresponding regions of interest within the diagnostic process. MAF = Missing Abnormality Finder; STARE: Spatio-Temporal Abnormal Region Extractor.

Radiology: Artificial Intelligence



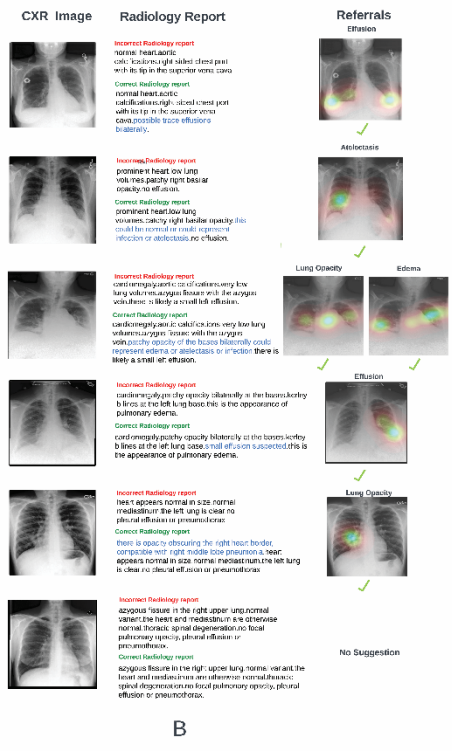


Figure 3: Examples of referrals generated by the system across different cases. **(A)** Referrals based on the random masking error dataset. **(B)** Referrals based on the uncertainty masking error dataset. Each subfigure is organized into three columns: the first column shows the actual CXR image, the second column presents the incorrect radiology report derived from simulated error data alongside its correct counterpart, and the third column displays the referrals generated by our system. Referrals are marked with a check mark if accepted by the radiologist or a cross mark if rejected. These acceptance and rejection marks are based on the original radiology reports rather than an independent review of the images by the radiologists. CXR = Chest X-ray.

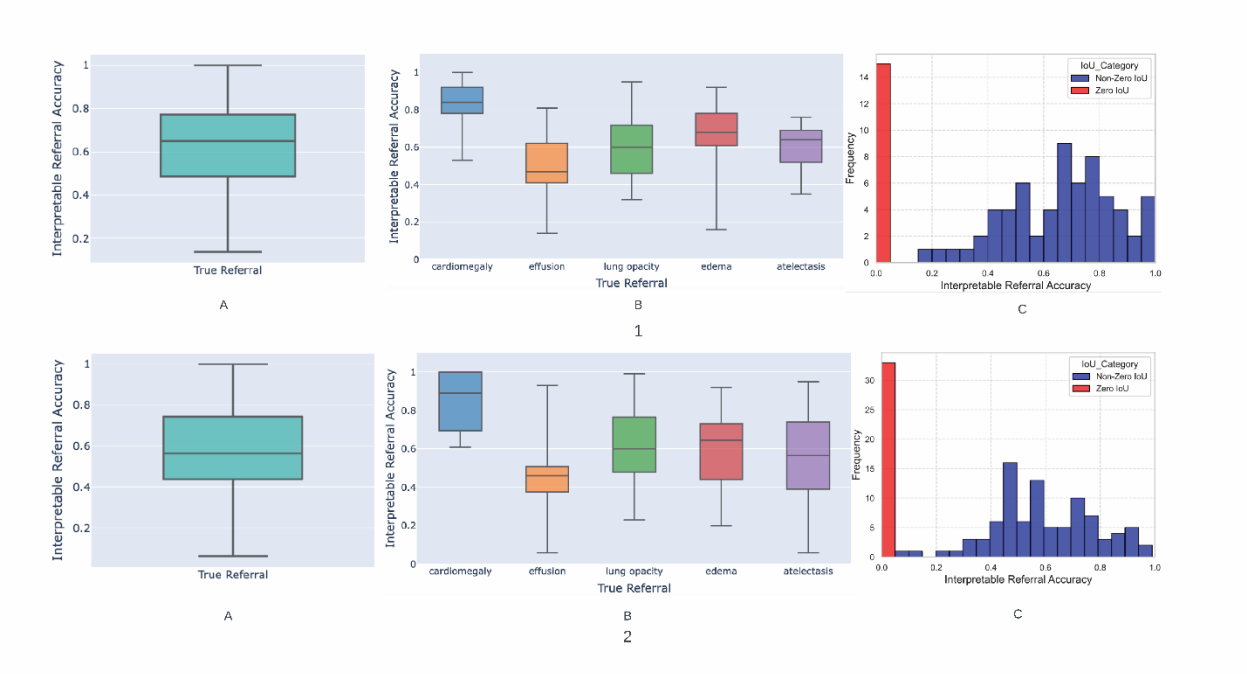


Figure 4: Comprehensive evaluation of CoRaX's referral-based interaction on two error datasets. Figure 4(1) presents results on the Random Masking-Based Error Dataset, while Figure 4(2) shows results on the Uncertainty Masking-Based Error Dataset. Each subfigure comprises three subplots (A, B, C). Plot A illustrates the distribution of Interpretable Referral Accuracy scores for all True Referrals (TR). Plot B displays the distribution of Interpretable Referral Accuracy scores for True Referrals (TR) corresponding to each abnormality. Plot C depicts the overall Interpretable Referral Accuracy of referrals (TR+FR) in the Referral-based Interaction.

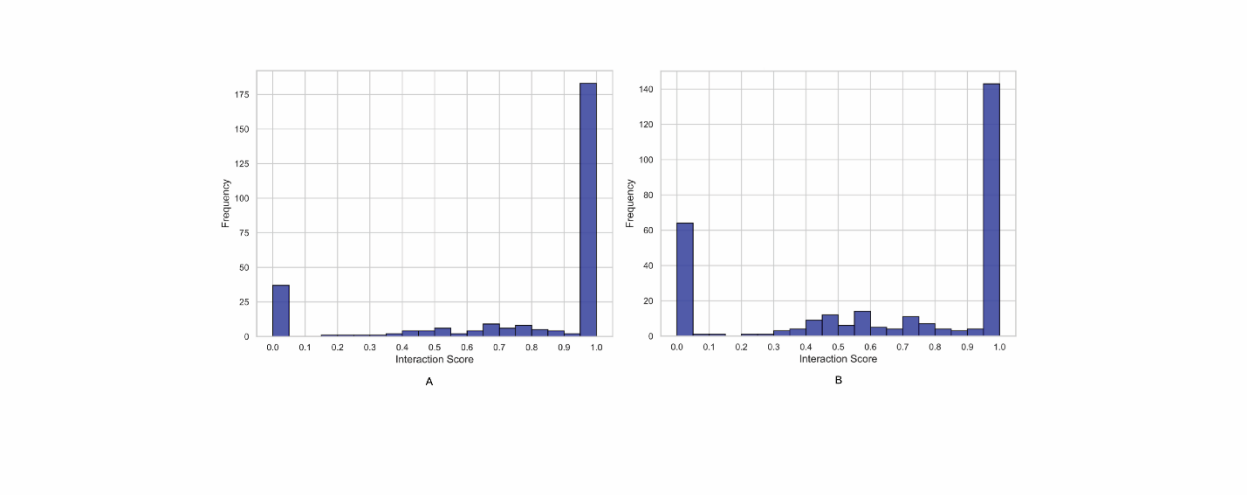


Figure 5: Histogram of Interaction score, reflecting the system's comprehensive performance considering the collaboration between the radiologist and CoRaX. Figure 5(A) illustrates the system's performance using the random masking-based error dataset, while Figure 5(B) displays results from the uncertainty masking-based error dataset. distribution function of the Interaction score, reflecting the system's comprehensive performance considering the collaboration between the radiologist and CoRaX.

Table 1: Overview of the Error Introduction Methods Used in the Test Set to Evaluate CoRaX's Performance

Abnormality	Error Based On Uncertainty Percentage (Error Cases/ Total Cases) %	Error Based On Random Masking Percentage (Error Cases/ Total Cases) %
Cardiomegaly	6.1(4/65)	15.3(10/65)
Pleural Effusion	44.6(29/65)	23.0(15/65)
Atelectasis	70.3 (38/54)	42.0 (23/54)
Lung Opacity	44.6(42/94)	27.6(26/94)
Edema	61.1(33/54)	35.1(19/54)

Note.—The dataset comprises 271 chest radiographs (CXR) with errors deliberately introduced to simulate real-world diagnostic challenges. Errors were introduced in two ways: (1) Random Masking of Abnormalities, where cases were altered without specific criteria to simulate general perceptual errors, and (2) Masking Based on Uncertain Abnormalities, where cases were adjusted according to uncertainty labels to reflect real-world diagnostic difficulties. The table presents the percentage of error cases for various abnormalities under these methods.

Table 2: Detailed Analysis of CoRaX's Performance in Identifying Visual Misses for Different Abnormality types in the Random Masking-Based Error Dataset

Abnormality	True Referrals (TR)	False Deferral (FD)	Perceptual Error Correction Rate (PECR)(%)	False Referral (FR)	True Deferral (TD)	Over-diagnosis Error Rate (ODER) (%)
Cardiomegaly	10	0	100.0(10/10) [72.2,100.0]	2	261	0.7(2/263) [0,1.9]
Edema	14	5	73.6(14/19) [50.0,93.9]	2	252	0.8(2/254) [0,1.9]
Atelectasis	14	9	60.8(14/23) [39.4,80.0]	2	248	0.8(2/250) [0.0,1.9]
Pleural Effusion	10	5	66.6(10/15) [44.1,90.2]	2	256	0.8(2/258)[0,1.9]
Lung Opacity	23	3	88.4(23/26) [74.1,100.0]	1	245	0.4(1/246)[0.0,1.5]

The table includes metrics for Perceptual Error Correction Rate (PECR) and Over-diagnosis Error Rate (ODER).

Table 3: Detailed Analysis of CoRaX's Performance in Identifying Visual Misses for Different Abnormality types in the Uncertainty Masking-Based Error Dataset

Abnormality	True Referrals (TR)	False Deferral (FD)	Perceptual Error Correction Rate (PECR)(%)	False Referral (FR)	True Deferral (TD)	Over-diagnosis Error Rate (ODER) (%)
Cardiomegaly	4	0	100.0(4/4) [51.0,100.0]	3	267	1.1(3/270) [0.3,3.7]
Edema	26	7	78.7(26/33) [75.7,100.0]	0	238	0.0(0/238) [0.0,2.0]
Atelectasis	28	10	73.6(28/38) [56.4,83.7]	2	233	0.8(2/235) [0.0,2.0]
Pleural Effusion	17	12	58.6(17/29) [55.5,89.2]	4	242	1.6(4/246) [0.0,2.0]
Lung Opacity	40	2	95.2(40/42) [85.2,100.0]	1	229	0.4(1/230) [0.0,1.7]

The table includes metrics for Perceptual Error Correction Rate (PECR) and Over-diagnosis Error Rate (ODER)

Table 4: System Performance on the Random Masking-Based Error Dataset, Showing the Frequency of Referral and Deferral Decisions Along with the Correctness of Each Decision Type

Interaction	Correct (%)	Incorrect (%)
Deferral-based	63.5(178/280)	7.8(22/280)
Referral-based	25.3(71/280)	3.2(9/280)

Table 5: System Performance on the Uncertainty Masking-Based Error Dataset, Showing the Frequency of Referral and Deferral Decisions Along with the Correctness of Each Decision Type

Interaction	Correct (%)	Incorrect (%)
Deferral-based	47.8(142/297)	10.4(31/297)
Referral-based	38.7(115/297)	3.3(10/297)

Supplementary material:

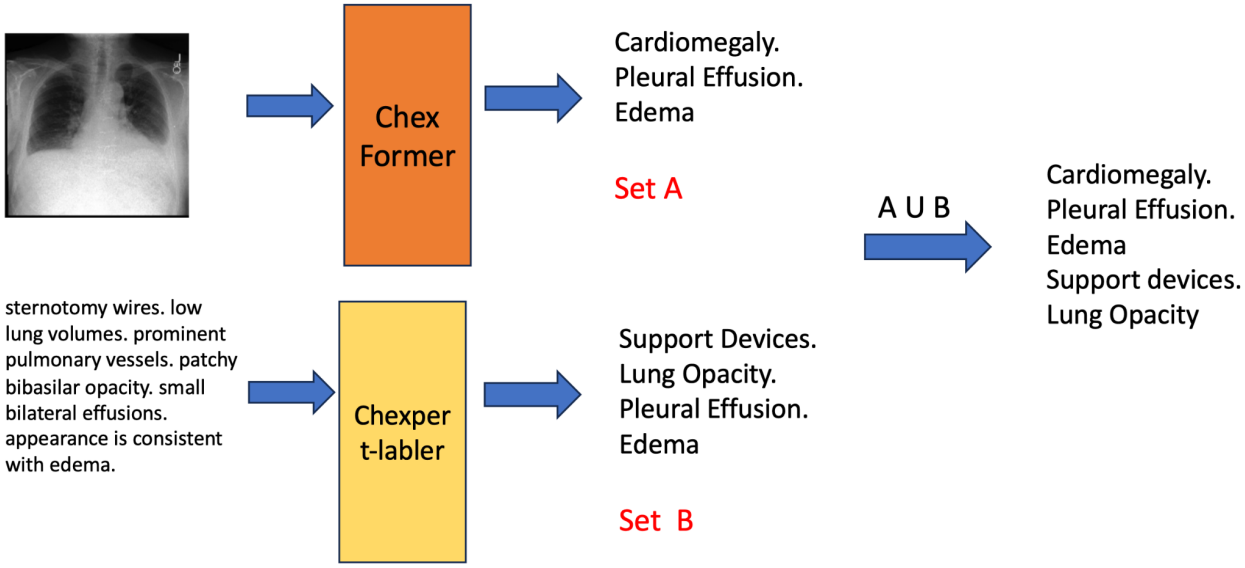
Detailed Methodology:

Missed Abnormality Finder: This module plays a pivotal role in summarizing and supplementing any overlooked abnormalities identified in the actual radiology report. In the initial phase of our system, MAF combines the CXR image and the actual radiology report to generate a summarized report with appended missing identified abnormalities.

Serving as the primary engine in our proposed system, MAF is created by combining the functionalities of the Chexpert Labeler and Chexformer.

The Chexformer, functioning as a multilabel classifier and the central component of the proposed system, takes the CXR image as input and produces multiple labels corresponding to the image (referred to as set A in supplementary figure 6). Concurrently, the actual radiology report undergoes processing by the Chexpert Labeler, an NLP tool designed for radiology report summarization. The Chexpert-Labeler extracts observations from free-text radiology reports and captures uncertainties present in the reports using an uncertainty label. It converts the detailed radiology report into 14 pathologies, defining three labels for each pathology as positive (1), negative (0), or uncertain (-1), and leaving blank for pathology if the doctor does not mention it in the report. While condensing the actual report into a summarized report, we consider all pathologies in the summarized report with -1 and 1 values because 0 indicates that the radiologist has not dismissed the

abnormality or has determined that there is no abnormality. Throughout the manuscript, the same principle is applied to the summarized radiology report. The labeler condenses the report into predefined labels, termed set B. The union of these two sets results in the final output – a summarized and corrected report.



Supplementary Figure 6: Design Blueprint for CoRaX's MAF Module, Comprising Two Key Components - Chexformer, a Multilabel Classifier, and the Chexpert Labeler. Chexformer performs multilabel classification on the provided CXR image, while Chexpert summarizes the actual report. The final output is the union of sets A and B, representing the corrected and summarized report. MAF = Missing Abnormality Finder

Chexformer: It serves as the fundamental engine of our proposed system—a multilabel transformer classifier designed to predict multiple labels corresponding to a given CXR image. This classifier, known as Chexformer, operates on a multi-label classification framework employing a Transformer encoder. Pretrained on the CheXpert dataset, Chexformer demonstrates proficiency in multilabel classification tasks.

The Transformer[1] plays a pivotal role as an architecture that captures essential relationships within input data using multi-head self-attention layers. Illustrated in

Supplementary Figure 7, Chexformer employs the Transformer encoder, utilizing label embeddings and image features extracted by a convolutional neural network during training to produce target labels. It learns the interaction between image features and label embeddings, modeling the joint embedding space or multimodal interactions.

In multi-label classification, the primary goal is to predict multiple labels denoted as $\{y_1, y_2, \dots, y_l\}$ where each y_i takes values of 0 or 1, given the chest X-ray image represented by x.

Let $H = \{z_1, \dots, z_{h\times w}, \tilde{l}_1, \dots, \tilde{l}_l\}$ represent the set of embeddings input to the Transformer encoder. In Transformers, the significance or weight of an embedding $h_j \in H$ concerning $h_i \in H$ is determined through "self-attention." The attention weight, denoted as α_{ij}^t between embedding i and j, is computed by first calculating a normalized scalar attention coefficient α_{ij}. After computing α_{ij} for all pairs of i and j, each h_i is updated to h_i' using a weighted sum of all embeddings, followed by a nonlinear ReLU layer.

$$\alpha_{ij} = softmax\left((W^q h_i)^\top (W^k h_j)/\sqrt{d}\right) \quad \text{- (1)}$$

$$\underline{h}_i = \sum_{j=1}^{M} \alpha_{ij}\, W^v h_j$$

$$h_i' = ReLU\left(\underline{h}_i W^r + b_1\right) W^o + b_2.$$

In this context, W^k represents the key weight matrix, W^q is the query weight matrix, W^v stands for the value weight matrix, W^r and W^o denote transformation matrices, and b_1 and b_2 are bias vectors. The outlined update procedure can be iterated for L layers, where the updated embeddings h_i' serve as input to the successive Transformer encoder layer. It's essential to note that the learned weight matrices $\{W^k, W^q, W^v, W^r, W^o\} \in R^{d\times d}$ are not shared across layers. The final output of the Transformer encoder, following L layers, is denoted as $H' = \{z_1', \dots, z_{h\times w}', l_1', \dots, l_l'\}$.

Image Feature Embeddings: For an input image $x \in R^{H \times W \times 3}$, the feature extractor generates a tensor $Z \in R^{h \times w \times d}$, where h, w, and d denote the output height, width, and channels, respectively. Each vector $z_i \in R^d$ from Z, where i varies from 1 to P (where $P = h \times w$), can be regarded as representing a subregion that corresponds to patches in the original image space.

Label Embeddings: For each image, we obtain a collection of label embeddings, denoted as $L = \{l_1, l_2, \ldots, l_l\}$, where each $l_i \in R^d$ represents one of the possible labels in y. These label embeddings are acquired through learning from an embedding layer with dimensions $d \times l$.

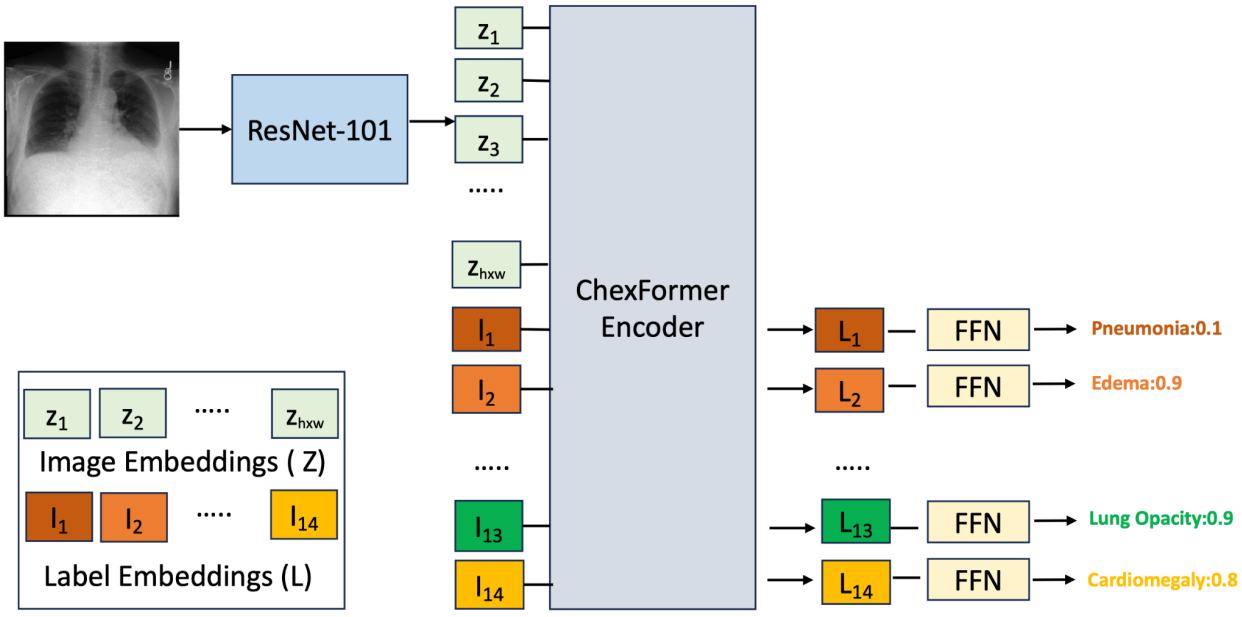
Image Feature Extractor: We employed ResNet-101[2] as a feature extractor, pre-trained on ImageNet[3]. The ResNet-101 output dimension is 2048, and we designated our embedding size, denoted as d, to be 2048. During Chexformer training, we resized the images to 640x640, performed random cropping to 576x576, and included some horizontal flips to introduce diversity in the training data. Conversely, testing images undergo center cropping. The ResNet-101 model generates an 18×18×d tensor as output, resulting in a total of 324 feature embedding vectors, represented as $z_i \in R^d$.

Transformer Encoder: In order to allow a particular embedding to attend to multiple other embeddings (or multiple groups), Chexformer uses 4 attention heads [1]. We use a L =3 layer Transformer with a residual layer [2] around each embedding update and layer norm [4].

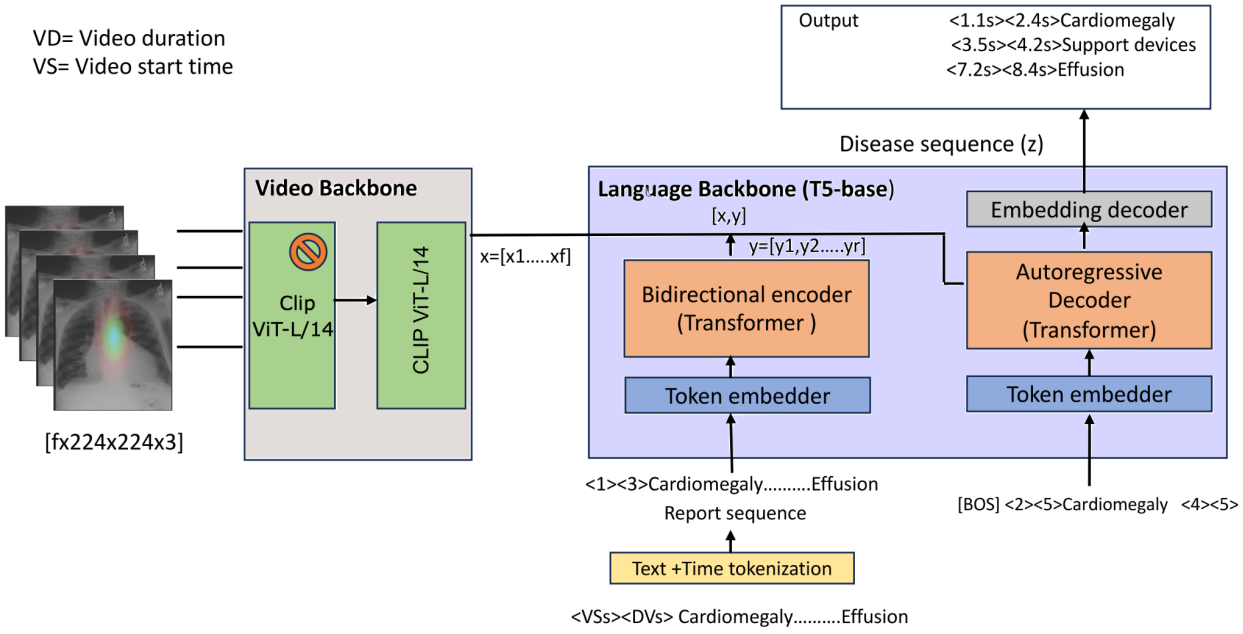
Training:ChexFormer undergoes training on the Chexpert dataset, employing Adam [5] as the optimizer with betas set to (0.9, 0.999) and zero weight decay. The training process utilizes a batch size of 16, a learning rate of 10^-5, and incorporates dropout [6] with a probability of p = 0.1 for regularization. ChexFormer is trained on a subset of approximately

4000 images from the Chexpert dataset, with an additional 1000 images utilized for testing ChexFormer. During training, ChexFormer takes label embeddings and image features as input, learning to predict labels. However, during the inference phase, it takes image features as input and outputs labels for the CXR image.

As previously mentioned, ChexFormer undergoes training on the CheXpert dataset. The dataset encompasses 14 labels, with each label capable of assuming values of 0, 1, or -1, representing no abnormality, confirmed abnormality, and uncertainty for the abnormality, respectively. In the training process of ChexFormer, instances of -1 are replaced with 1. This adjustment is made because our ultimate goal is to offer recommendations to radiologists regarding potentially overlooked abnormalities.



Supplementary Figure 7: depicts an illustration of Chexformer, a pretrained multilabel classifier on the CheXpert dataset. In the training phase, it receives the CXR image and label embedding as input, producing scores for each label while learning the interaction between labels and image features.



Supplementary Figure 8: STARE module overview: A sequence to sequence model which takes video features and summarized radiology report with appended time tokens as input and outputs the abnormality sequence with temporal grounding.

Spatio-Temporal Abnormal Region Extractor: The core of our system features the Spatio-Temporal Abnormal Region Extractor (STARE) module, tasked with predicting the temporal alignment of each abnormality outlined in the condensed radiology report and then use temporal alignment to find the corresponding region of interest for each abnormality. The architecture of this module is delineated in Supplementary Figure 8, consisting of two essential components: the Video Backbone and the Language Backbone. Our system design is inspired by deep video captioning in computer vision[7,8]. During the inference phase, the condensed radiology report is enriched with start and end times. For

text radiology reports, the start time is initialized at 0 seconds, while the end time corresponds to the video duration. In speech transcription scenarios, the start time signifies the commencement of the radiologist's report depiction, and the end time signals the radiologist's conclusion.

STARE processes both the frames of the radiologist's eye gaze fixation video $x = \{x_i\}_{i=1}^{F}$ and the time-token-appended condensed and corrected radiology report $y = \{y_j\}_{j=1}^{S}$, where y denotes the abnormality or label from the CheXpert dataset. The model's output is a abnormality sequence $z = \{z_k\}_{k=1}^{L}$, encapsulating each abnormality's textual description and timestamps indicating the temporal abnormality locations in the video. The term "condensed and corrected radiology report" refers to the output of the Missing Abnormality Finder module, representing the union of the CheXpert labeler output and the labels predicted by CheXformer. However, during training, STARE independently undergoes training on the REFLACX and EGD-CXR datasets.

Video Backbone

The Video Backbone plays a pivotal role in extracting features from the input video. It comprises a spatial encoder followed by a temporal encoder, operating on a sequence of 'f' frames. Utilizing a pre-trained CLIP ViT-L/14[10,11] as the spatial encoder, we extract individual frame features, considering the spatial characteristics of each frame in the video. The frame of each video resized to 224x224 before extracting the features. The input set consists of videos with dimensions '$f \times h \times w \times c$', where '$h$', '$w$', and '$c$' represents the height, width, and number of channels of each frame. The spatial encoder processes each frame independently, and we maintain the spatial backbone as frozen to minimize computational costs and parameter count in the overall model.

The spatial encoder generates a two-dimensional array, with the first dimension representing the number of frames and the second representing the embedding dimension. Although each video may have a varying number of frames, we limit our consideration to the features of the first 100 frames. To accommodate videos with fewer than 100 frames, we pad the feature extraction output from Resnet with zeros.

For the temporal encoder, we employ a pre-trained CLIP ViT-L/14[10,11] transformer to produce contextualized embeddings, contributing to the comprehensive feature representation of the input video.

Language Backbone

Our language backbone[59] is built on the T5[12] , employing an encoder-decoder architecture. We initialized both the text encoder and decoder with the t5-base model, which underwent pretraining on web text corpora with a denoising loss.

Text and Time tokenization

We utilize the SentencePiece tokenizer[11] with a vocabulary size of $V = 32{,}128$. Our approach involves initial text tokenization, and to augment this process, we incorporate two extra time tokens, bringing the total to V + 2 tokens. Throughout the training, these time tokens represent the initiation and conclusion times when the radiologist begins and concludes the depiction of the radiology report while examining the CXR image on the screen. The time tokenization process adheres to the equation detailed below.

$$tt = \left| \frac{(ts \times N)}{D} \right| \qquad (1)$$

In equation 1, "tt" denotes the time token, "ts" represents the timestep (indicating the start or end time step), "N" signifies the quantized bin with a specified value of 100 (N=bins), and "D" corresponds to the video duration.

Text Encoder

It accepts a report sequence as input, where the report sequence comprises 'r' tokens denoted as 'y' belonging to the set $y \in \{1, \ldots, V + N\}^r$. Here, 'v' represents the vocabulary size of text, 'n' is the size of time tokens (here $n = 2$) , and 'r' stands for the total number of tokens in the report sequence. The text encoder includes an embedding layer responsible for independently embedding each token, producing a semantic embedding of size 'rxd'. Subsequently, a transformer encoder calculates contextualized embeddings of size 'rxd', with 'd' representing the hidden dimension.

Text Decoder

Comprising a transformer decoder and an embedding layer, the system generates an abnormality sequence with associated temporal grounding, referred to as the abnormality sequence. Each abnormality k is characterized by a text segment, a start time and an end time. We first construct for each event k a sequence by concatenating its start time token t(startk) , its end time token t(endk), and its text tokens $[z_{k1}, \ldots, z_{kl_k}]$. Finally, the event sequence is obtained by prepending and appending a BOS and EOS tokens to indicate the start and the end of the sequence, respectively, i.e. $z = [BOS, t_{start\ 1}, t_{end\ 1}, z_{1_1}, \ldots, z_{1_{l_1}}, t_{start\ 2}, \ldots, EOS]$.

The transformer decoder, functioning causally, employs cross-attention with the encoder output, formed by concatenating visual and encoder transformer embeddings (x_t and y_t), along with all tokens generated earlier. Simultaneously, it performs self-attention across the entire set of previously generated tokens. The text decoder produces the event sequence z by utilizing an embedding decoder, which is applied on top of the transformer text decoder. This decoder predicts the probability distribution over the joint vocabulary of text and time tokens, enabling the model to anticipate the subsequent token in the report sequence.

Region extraction

The process involves using the predicted time steps (start and end time) for the missing abnormality to extract frames within this interval. The mean image is then calculated using all these extracted frames, providing a comprehensive representation of the overall intensity in the critical region. Similarly, frames are extracted using the ground truth time steps for the predicted missing abnormality, and the mean image is calculated from these frames, consolidating multiple extracted frames into a single image. But we also provide an option in the system if the user wants to have a more detailed look then the user can also pool all the fixation points between the extracted time step into a single heatmap called static heatmap without taking a mean image. The comparison between the predicted mean image and the ground truth mean image is then performed by calculating the IoU.

Fine Tuning

We utilized the pretrained model from vid2seq[7], specifically trained on the ActivityNet Captions dataset[13], which comprises approximately 20,000 untrimmed videos depicting diverse human activities. Each video is accompanied by transcribed speech sentences and timestamps, establishing a temporal connection to events. Given the limited availability of fixation videos and corresponding transcriptions in the medical domain, leveraging this pretrained model enables our system to understand long-term relationships among different speech segments.

During the fine-tuning stage, the model undergoes refinement to predict the temporal grounding of each abnormality in the summarized radiology report obtained through the CheXpert labeler. The fine-tuning objective is rooted in the maximum likelihood objective, as further detailed in this context[7]. The primary goal of this Spatio-Temporal Abnormal Region Extractor (STARE) module is to understand human cognition during decision-making in abnormality diagnosis. In simpler terms, this module learns what radiologists focus on when making decisions based on CXR images.

Set Difference operation to generate the referral: To identify missing abnormalities in the actual radiology report, we compare it with the output of the Chexpert-Labeler [14] module. In the main manuscript Figure 2, Set A represents the actual summarized radiology report, and Set B represents the output of the STARE module, excluding timesteps for the analysis of the set difference. The set difference reveals missing abnormalities in the radiology report. Set B has a cardinality greater than or equal to that of set A, aligning with our system's focus on correcting perceptual errors or finding missing diagnoses.

The set difference result reveals missing abnormalities, and we extract the corresponding timesteps from the STARE module output. Using these timestamps, we identify fixation points between them, representing frames in the video. If this time interval spans multiple frames, we compute the mean of the image frames and merge them into a single image, referred to as the region of interest for the missing abnormality. But we also provide an option in the system if the user wants to have a more detailed look, then the user can also pool all the fixation points between the extracted time step into a single heatmap called a static heatmap without taking a mean image. This consolidated image and the missed abnormality serve as the referral produced by our system.

Data Preprocessing

The speech transcription data obtained from EGD-CXR and REFLACX contains detailed radiology reports with word alignments for CXR images. By merging the transcriptions from REFLCAX and EGD-CXR, we generated a final JSON file that includes comprehensive reports and associated timesteps. This compilation is crucial for training the temporal grounding predictor module.

In our preprocessing phase, we summarize the real radiology reports to acquire both the ground truth and input report necessary for training the Spatio-Temporal Abnormal Region Extractor (STARE) module. When condensing the radiology reports, we prioritize essential abnormalities outlined in the CheXpert labeler. Our approach to preprocessing aims to

avoid converting the entire radiology report into isolated labels, instead focusing on ensuring that the model grasps the fundamental aspects of chest X-ray anatomy, including spatial relations like "right" and "left" lungs. Radiology reports typically comprise multiple sentences separated by periods. To address this structure, our preprocessing methodology involves extracting each sentence. During this extraction process, we meticulously scrutinize each sentence for phrases corresponding to abnormalities in the CheXpert labeler. If a match is found, we substitute the sentence with the relevant abnormality; otherwise, it remains unchanged.

The extraction of timesteps for each abnormality or unchanged sentence relied on the speech transcription associated with each radiology report. This resulting file served as the ground truth during the model fine-tuning, encompassing fixation heatmap videos and summarized radiology reports featuring only start and end timestamps for the entire report as input. This ensured a robust foundation for training. The core objective of the Spatio-Temporal Abnormal Region Extractor (STARE) module is to comprehend human cognition by predicting the timestamps associated with each intention. It's worth noting that our focus does not extend to conducting dense video captioning in this context.

Random-Masking-Based Error Dataset:

Inexperienced radiologists are more prone to missing cases due to their limited knowledge, whereas experienced radiologists typically have a lower rate of missed cases. However, experienced radiologists may still miss cases if their fixation time is shorter or if they do not pay sufficient attention to abnormal regions while diagnosing specific abnormalities. Limited focus or attention may be a primary reason for visual misses among experienced radiologists, but other factors such as fatigue or poor lighting could also contribute.

To cover all potential scenarios, we chose to randomly alter abnormalities in the test set. These alterations, considered anomalies, are intentionally kept at a low percentage in the dataset. The primary aim of this study is to identify perceptual errors and rectify them. When we mention "altered," it means that if a specific case originally includes a particular

abnormality, we either mask it or alter it with negation. Both scenarios fall under the category of perceptual errors resulting from recognition or decision-making. Consequently, specific abnormalities listed in Table 1 have been randomly modified in the test data at varying percentages, as indicated below.

Since the proposed system takes the summarized report as input, whether we mask the abnormality or alter it with negation does not affect the outcome. In both cases, the abnormality will not be included in the summarized report.

Referral Evaluation Metrics:

To evaluate the accuracy of referrals in identifying missed abnormalities, we assessed how many times such abnormalities are correctly identified and corrected. This assessment includes calculating True Referral (TR), False Referral (FR), False Deferral (FD), and True Deferral (TD) for each abnormality.

True Referral (TR) refers to the number of abnormalities that were missed by radiologists but correctly identified by CoRaX**.** It also represents the number of referrals accepted by radiologists.

False Referral (FR) refers to the number of abnormalities that CoRaX incorrectly over-diagnoses (flags as abnormal when they are not). It also represents the number of referrals rejected by radiologists.

False Deferral (FD) refers to the number of abnormalities that were missed by radiologists, and the system also fails to identify these abnormalities. In this case, both the radiologist and the system overlook the abnormality, leading to the incorrect decision of not making a referral.

True Deferral (TD) refers to the number of abnormalities for which the system correctly decides not to make a referral. This occurs when the abnormality is identified as non-critical or has already been addressed by the radiologist.

Using TR, FR, FD, and TD, we subsequently compute the Perceptual Error Correction Rate (PECR) and Overdiagnosis Error Rate (ODER), providing a singular metric for understanding the model's performance in identifying errors.

Perceptual Error Correction Rate (PECR): The metric measures the system's effectiveness in rectifying perceptual errors for each abnormality, with individual PECR values outlined in Table 1. This metric, resembling the system's recall, is defined as follows:

$$PECR(\%) = \frac{(TR)}{TR+FD} \quad (3)$$

Over-diagnosis Error Rate (ODER): This metric measures how frequently the model over-diagnoses abnormalities, indicating the proposed system's error rate. It resembles the system's FPR(False positive Rate):

$$ODER(\%) = \frac{(FR)}{FR+TD} \quad (4)$$

Evaluation of Referrals based on the region of interest: The system's ability to identify missing abnormalities and highlight their regions of interest is assessed by measuring the overlap between the true and predicted regions, utilizing the Intersection over Union (IoU) metric.

Higher IoU scores indicate better accuracy of the highlighted region of interest, indicating reduced confusion for radiologists and offering insights into potentially saving clinical time. By effectively directing radiologists to the correct regions of interest for missed abnormalities, the system minimizes the need for them to initiate their search anew.

$$Interpretable\ Referral\ Accuracy\ (IRA)(\%)\ = I(Referral\ is\ Accepted) * IoU * 100 \quad (5)$$

Where I is an indicator function whose value is 1 if the referral is accepted.

Evaluation of overall interaction and diagnostic accuracy: We introduce the Interaction score metric, which serves as an indicator of the diagnostic accuracy of each interaction between the system and the radiologist. An Interaction score of 1 indicates a fully beneficial interaction, representing diagnostic aid with no confusion. The Interaction score is calculated based on the case level.

Interactions between the system and the radiologist can be classified into two categories: referral-based and non-referral-based. We consider the interaction as non-referral or Deferral if there is not any referral for a particular case. For non-referral interactions, a score of 1 is assigned when the system refrains from making any referral for a particular case and its decision aligns with the correct diagnosis, ensuring no perceptual errors are overlooked. Conversely, a score of 0 is assigned when the system makes an incorrect deferral, indicating 0% diagnostic accuracy and no assistance to the radiologist. Referral-based interactions occur when the system makes one or more referrals for a particular case. Each referral for a case is counted as a separate referral-based interaction. For example, if a case involves three referrals, it contributes three referral-based interactions. This ensures that the total number of interactions accounts for all system-generated referrals, which may exceed the number of cases. In referral-based interactions, an Interaction score of 1 represents both 100% diagnostic accuracy and 100% spatial precision in identifying abnormal regions. We utilize the IoU score to quantify the spatial precision of the referrals.

Bootstrap for CI Calculation: The bootstrap method was used to calculate the 95% confidence intervals (CIs) for the PECR, ODER, Interpretable Referral Accuracy, and Interaction Score. This method is a statistical resampling technique that estimates the distribution of a metric by repeatedly sampling from the observed data with replacement.

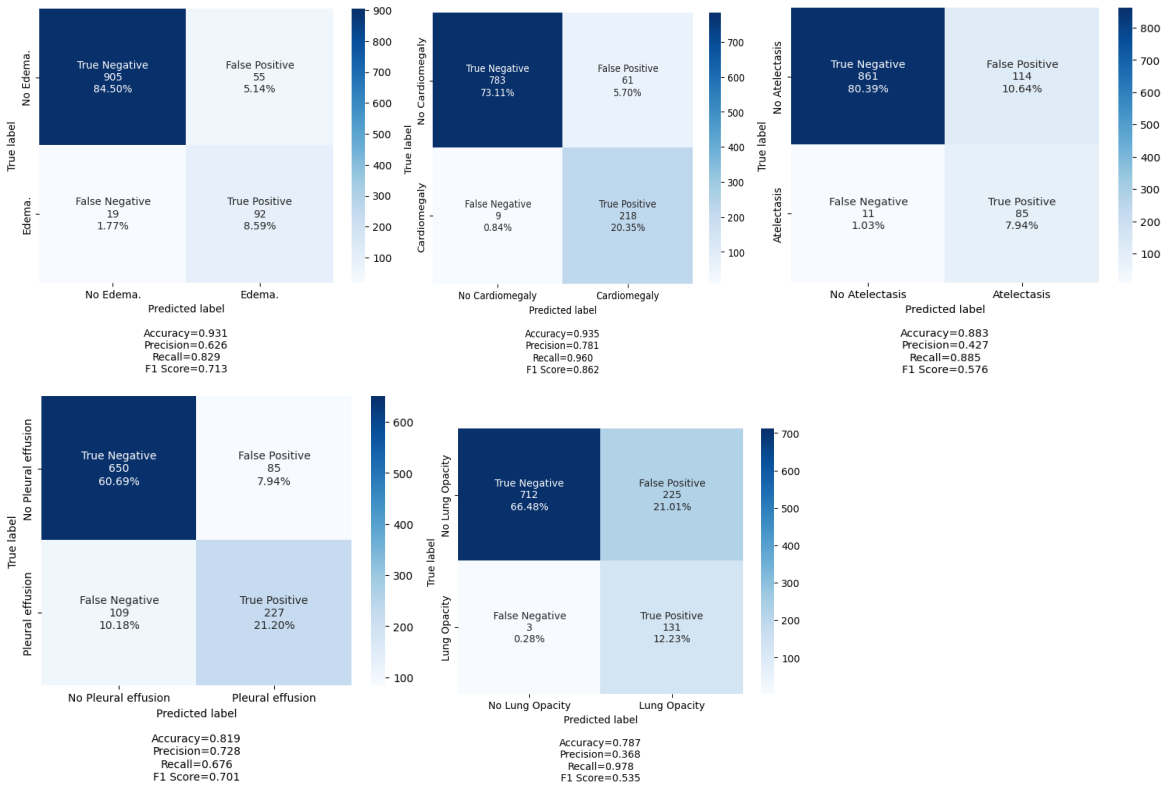
To apply the bootstrap method, the original dataset was used as the basis for resampling. A total of 1000 bootstrap samples were generated, where each bootstrap sample was the same size as the original dataset but contained values randomly drawn with replacement. This process ensures that some observations may appear multiple times in a single sample, while others may be omitted. For each of these 1000 resampled datasets, the desired metric (e.g., PECR, ODER, IRA, and Interaction score) was computed. This resulted in a distribution of 1000 metric values based on the resampled data.

The 95% CI for each metric was then determined by sorting the 1000 bootstrap estimates in ascending order and identifying the values corresponding to the 2.5th and 97.5th percentiles. These percentile values represent the lower and upper bounds of the confidence interval, respectively. By employing the bootstrap method, we ensured that the confidence intervals for all metrics were calculated without relying on assumptions about the underlying data distribution, making the results both reliable and interpretable.

This resampling approach provides robust estimates of variability without assuming a specific data distribution.

However, in cases where a metric achieved a **perfect score** (e.g., PECR = 100% for cardiomegaly), the bootstrap method produced **degenerate intervals** such as [100.0%, 100.0%], which fail to capture the underlying uncertainty. To address this, we used the **Wilson score interval**, a more reliable method for estimating confidence intervals for proportions near boundary values (0% or 100%). The Wilson interval was computed using a significance level of $\alpha = 0.05$, corresponding to a 95% confidence level. This method was specifically applied for calculating the CI for PECR in the cardiomegaly subgroup.

Results of the Chexformer on the test set of ChexFormer dataset



Supplementary Figure 9: ChexFormer's performance evaluated on the Chexpert dataset. The confusion matrix illustrates ChexFormer's classification performance for each abnormality.

Results of the STARE module on the test set of REFLACX and EGD-CXR dataset

Abnormality	Mean-IoU (EGD-CXR)	Mean-IoU (REFLACX)
Cardiomegaly	0.58 [0.53,0.62]	0.56 [0.50,0.61]
Atelectasis	0.53 [0.49,0.58]	0.60 [0.57,0.62]
Edema	0.63 [0.59,0.67]	0.63 [0.60,0.67]
Pleural Effusion	0.51 [0.48,0.54]	0.49 [0.47,0.50]
Lung Opacity	0.59 [0.54,0.63]	0.61 [0.59,0.64]

Supplementary Table 4. Mean IoU scores for the predicted ROIs corresponding to each abnormality. The table summarizes the performance of the STARE module across two datasets (EGD-CXR and REFLACX). 95% CIs were computed using bootstrapping. ROI = Region of Interest; STARE: Spatio-Temporal Abnormal Region Extractor

Abnormality specific IRA Scores

Abnormality	**Random Masking-Based Error Dataset (IRA) (%)**	**Uncertainty Masking-Based Error Dataset (IRA) (%)**
Cardiomegaly	83.0 [74.0, 90.0]	84.0 [69.0, 100.0]
Atelectasis	60.0 [54.0, 66.0]	56.0 [48.0, 65.0]
Pleural Effusion	50.0 [38.0, 63.0]	45.0 [38.0, 54.0]
Lung Opacity	61.0 [55.0, 69.0]	62.0 [57.0, 68.0]
Edema	64.0 [54.0, 73.0]	60.0 [53.0, 67.0]

Supplementary Table 5. Mean IRA scores for each abnormality across both simulated error datasets. This table summarizes the performance of the CoRaX system in predicting ROIs for missed abnormalities, stratified by abnormality type. IRA = Interpretable Referral Accuracy.

Supplementary Material References:

[1]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.

[2]He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[3]Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[4]Ba, J.L., Kiros, J.R. and Hinton, G.E., 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

[5]Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[6]Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), pp.1929-1958.

[7]Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J. and Schmid, C., 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10714-10726)

[8]Iashin, V. and Rahtu, E., 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 958-959).

[9]Wang, Z., Wu, Z., Agarwal, D. and Sun, J., 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

[10]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[11]Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

[12]Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), pp.5485-5551.

[13]Krishna, R., Hata, K., Ren, F., Fei-Fei, L. and Carlos Niebles, J., 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 706-715).

[14] Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. InProceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 590-597). doi: 10.1609/aaai.v33i01.3301590

RSNA

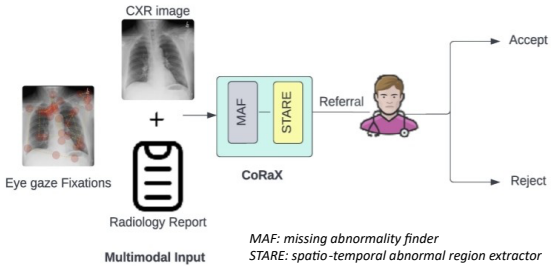
Collaborative Integration of AI and Human Expertise to Improve Detection of Chest Radiograph Abnormalities

Key Result
The proposed system, CoRaX, demonstrated potential to aid radiologists to aid radiologists in detection of abnormalities on chest radiographs by identifying and correcting perceptual errors.

Datasets:
- EGD-CXR dataset (n = 1,071 CXRs)
- REFLACX dataset (n = 2,440 CXRs)

Methods:
- The collaborative AI solution, named Collaborative Radiology Expert (CoRaX), employs a large multimodal model to analyze image embeddings, eye gaze data, and radiology reports, aiming to rectify perceptual errors in CXR interpretation.
- The system's referral-making process, the quality of referrals, and its performance within collaborative diagnostic settings were evaluated.

Results:
- CoRaX corrected 21.3% and 34.6% of errors, defined as alterations in abnormalities, in the two simulated error datasets.
- In a collaborative diagnostic setting, the system had a non-zero interaction score in 85.7% and 78.4% of interactions for the error datasets, indicating its potential to aid radiologists in diagnostic decision-making.



Awasthi A et al. Published Online: July 16, 2025
https://doi.org/10.1148/ryai.240277