Evaluating the Diagnostic Performance of ChatGPT-4o Mini in the Classification of Chest X-Ray Pathologies

Swapna Vaja, BS[1]✉, Rahul Kumar, BS[2], Tejas C. Sekhar, BA[1], Louis Clarkson, MB BCh BAO[3], Nitin Chetla[4], Rahul Reddy[4], Shivam Patel[4], Kyle Sporn, MS[5], Phani Paladugu, MMSc, MS[6,7], Amar S. Vadhera, BS[7], Ahab G. Alnemri, MBA[8], Joshua Ong, MD[9], Ethan Waisberg, MB BCh BAO[10]✉, Mouayad Masalkhi, MB BCh BAO[11,12], Ram Jagadeesan, MS[13,14], Nasif Zaman, MS, PhD[15], Alireza Tavakkoli, PhD[15], Kunal Sukhija, MD[16]

[1] Rush Medical College, Chicago, Illinois, United States

[2] Department of Biochemistry and Molecular Biology, University of Miami Miller School of Medicine, Miami, Florida, United States

[3] School of Medicine, University of Cambridge, Cambridge, United Kingdom

[4] University of Virginia School of Medicine, Charlottesville, Virginia, United States

[5] Upstate Medical University Norton College of Medicine, Syracuse, New York, United States

[6] Brigham and Women’s Hospital, Boston, Massachusetts, United States

[7] Sidney Kimmel Medical College, Philadelphia, Pennsylvania, United States

[8] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States

[9] Michigan Medicine, Ann Arbor, Michigan, United States

[10] Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom

[11] College of Medicine, University of Dublin, Ireland

[12] Department of Electronic and Computer Engineering, University of Limerick, Ireland

[13] Cisco Artificial Intelligence Systems, Cisco Inc., San Jose, California, United States

[14] Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, United States

[15] Department of Computer Science, University of Nevada, Reno, Reno, Nevada, United States

[16] Kaweah Health Medical Center, Visalia, California, United States

Corresponding Author:

✉ Ethan Waisberg, MB BCh BAO

Department of Clinical Neurosciences, University of Cambridge,

Downing Street, Cambridge CB2 3EH, United Kingdom

Email: ew690@cam.ac.uk, cambridgemlgroup@gmail.com

ORCid: 0000-0001-8999-0212

Statements and Declarations:

Funding: The authors did not receive support from any organization for the submitted work.

Competing Interests: Ram Jagadeesan: Cisco (Employee, Stock). All other authors declare no conflicts of interest.

Ethics approval and consent to participate: Not applicable.

Consent for publication: The authors consent to publication of this manuscript.

Availability of data and materials: Datasets are available at:
https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/data

Authors' contributions:

Conceptualization: SV, RK, KS, EW

Methodology: RK, TCS, LC, KS, RJ

Software: RJ, NZ, AT

Validation: RK, KS, RJ

Formal analysis: RK, KS, PP

Investigation: RK, TCS, LC, NC, RR, SP, ASV

Resources: EW, JO, MM, RJ

Data Curation: RK, TCS, SV

Writing – Original Draft: RK, SV, TCS

Writing – Review & Editing: EW, JO, LC, KS, PP, RJ

Visualization: TCS, RR, SP, NC

Supervision: JO, EW, KS, KSuk

Project Administration: EW, JO, AG

Funding Acquisition: Not applicable

Acknowledgements
None.

Clinical trial number
Not applicable.

Abstract:

Chest X-ray (CXR) imaging is fundamental in diagnosing thoracic pathologies, yet its interpretation is highly dependent on radiological expertise. In this study, we evaluate the diagnostic performance of OpenAI's ChatGPT-4o mini in classifying 14 distinct CXR pathologies using the Kaggle train subset of the VinDr-CXR dataset, a publicly available CXR dataset from Vietnamese hospitals optimized for artificial intelligence (AI)-based medical image analysis. We queried the model using a standardized prompt and assessed its classification performance using accuracy, precision, recall, F1-scores, and a multi-class confusion matrix.

Our results reveal significant limitations in ChatGPT-4o mini’s ability to differentiate between pathologies, with an overall accuracy of 0.05 and macro-averaged precision, recall, and F1-scores of 0.28, 0.17, and 0.14, respectively. Several conditions, including aortic enlargement, interstitial lung disease (ILD), and pneumothorax (PTX), were entirely misclassified, yielding F1-scores of 0.00. We observed strong correlations between classification performance and dataset imbalances, as pathologies with higher support values demonstrated improved recall and F1-scores. Results obtained were used to generate a confusion matrix, which further highlights the model’s extreme misclassification rates—particularly for underrepresented classes—raising concerns about its generalizability.

Our findings underscore the challenges of applying large language models (LLMs) to direct CXR interpretation. We emphasize the need for improved training methodologies, dataset balancing, and alternative AI architectures to enhance diagnostic accuracy. While ChatGPT-4o mini may have potential for ongoing and future clinical applications, its current performance remains insufficient for independent diagnostic use, necessitating further refinement before integration into clinical workflows requiring continuous evaluation and monitoring.

Introduction:

Chest X-rays (CXR) are the most common type of imaging procedure used in clinical practice today, with radiologists primarily responsible for making correct diagnoses from CXR findings—including pneumonia, tuberculosis, lung cancers, and cardiomegaly [1]. However, CXR interpretation is affected by nuances of embedded human physiology, unique individual anatomy, patient positioning, and imaging artifacts. In response to the growing demand of radiological services and interpretation to meet the needs of a growing and aging population, highly efficient and effective diagnostic systems are ultimately needed to manage an increasing workload and clinical volume in parallel [2]. AI techniques, specifically deep learning, wields promise as a possible approach to enhance the clinical interpretability of CXRs. For example, deep learning architectures like convolutional neural networks (CNNs) have proven effective in pattern recognition and disease classification among large, expert-annotated datasets [3]. However, due to the current limitations in algorithmic design, AI models tend to struggle with performing multi-class classification of CXR abnormalities from images contained in the examined Kaggle train subset. Specifically, a major problem in AI analysis of CXR images is the heterogeneity of the imaging data, seeing as patient demographics, disease manifestations, imaging techniques, and image quality are critical factors that influence clinical interpretation and collectively pose challenges to model training and inference. Furthermore, the pervasive issue of class imbalance in medical datasets is often overlooked, with certain conditions—including focal lesions (e.g., lung nodules/masses) and non-lesional abnormalities (e.g., pleural effusion (PE), cardiomegaly, aortic enlargement, and PTX)—being underrepresented. This imbalance results in biased predictions and poor generalization to the patient population [4-6].

Current state-of-the-art CNN-based models have achieved superior results in binary classification tasks, such as normal and abnormal scan classification, but the challenge of multi-class classification still remains unresolved [7]. CNNs are specifically designed for image analysis and excel in tasks such as medical imaging and pattern recognition. In contrast, LLMs like ChatGPT-4o Mini are optimized for natural language processing (NLP) and are not inherently structured for direct image interpretation, though some multimodal AI models attempt to integrate both capabilities [ ]. Although LLMs are promising in clinical practice, especially with regard to medical text analysis and recommendation, their use in interpreting CXR images has not been thoroughly investigated. Our study aims to evaluate the application of LLMs as a diagnostic tool, specifically assessing their ability to interpret textual information from CXR reports while attempting to infer radiological findings without direct image analysis.

In this study, we examined the VinDr-CXR dataset, which contains 18,000 radiologist-annotated CXR scans covering 14 documented pathologies, to evaluate the diagnostic performance of ChatGPT-4o mini in the multi-class classification of CXR pathologies. To ensure a clinically relevant and statistically meaningful evaluation, we selected 14 pathologies based on the following criteria: (1) their documented prevalence in clinical practice, (2) their representation in the dataset to ensure sufficient sample size for model assessment, and (3) their relevance in AI-based diagnostic classification. Taken together, this subset of 14 pathologies comprised X cases, or Y% of the entire VinDr-CXR dataset.

Although most selected pathologies had a sufficient number of cases to support performance evaluation, pneumothorax (PTX) presented a notable limitation, with only 19 cases available in the dataset. Despite its low representation, PTX was included due to its clinical significance in emergency medicine and prior evidence of AI models struggling with its detection. However, its limited sample size may have impacted the model's ability to generalize to PTX cases, a consideration further explored in the discussion of classification performance and dataset imbalances.

The 14 pathologies selected include a combination of high-prevalence conditions (e.g., PE, lung opacity, PTX) and diagnostically challenging findings that AI models have historically struggled with (e.g., interstitial lung disease, aortic enlargement) [ ]. Pathologies with extremely low representation in the dataset, such as conditions with fewer than X images, were excluded to prevent artificially skewed performance metrics. A detailed numerical breakdown of the total dataset's pathology distribution and the subset composition used in this study is provided in Table X, following a dataset stratification approach to ensure transparency in case selection.

Subset selection was also guided by computational feasibility and financial constraints associated with API call costs and inference time. Given the structure of OpenAI's API, querying ChatGPT-4o mini for all 22 findings across a dataset of 18,000 images would have incurred substantial computational overhead and made large-scale evaluation impractical. Limiting the scope to 14 key pathologies allowed for a more controlled and reproducible evaluation while maintaining statistical rigor.

By consolidating clinically significant, well-represented, and diagnostically challenging pathologies, we aimed to characterize the current limitations of LLMs in CXR interpretation while ensuring that our findings contribute meaningfully to ongoing discussions on AI integration into medical imaging workflows.

Methods:

A structured evaluation was conducted with OpenAI's ChatGPT-4o mini API to determine its diagnostic accuracy in the classification of presence vs. absence of each of the 14 selected pathologies in provided CXRs from the Kaggle train subset of the VinDr-CXR dataset. The full VinDr-CXR dataset comprises

18,000 radiologist-labeled images, while the Kaggle train subset specifically used in this study contains 15,000 images. Pathologies outside this subset were excluded based on dominant pathology prevalence and relative frequency thresholds to ensure robust model evaluation. Specifically, conditions with fewer than X cases or those frequently co-occurring with other findings (leading to labeling ambiguities) were removed to prevent artificially skewed classification performance. This selection process prioritized pathologies with clear diagnostic boundaries and sufficient sample representation for meaningful analysis. Taken together, this subset of 14 pathologies comprised X cases, or Y% of the entire VinDr-CXR dataset.

To ensure the dataset was in a suitably structured format for LLM evaluation, we performed our own pre-processing evaluation that involved detailing the presence and frequency of each pathology. Within this subset of 14 pathologies, we identified 568 images with a "dominant pathology," defined as cases where a single pathology occurred at least three times more frequently than any other pathology. To maintain dataset integrity and ensure comparability in model evaluation, we did not modify, collect, or relabel any scans.

To assess whether ChatGPT-4o mini could accurately identify pathologies, we designed a standardized diagnostic prompt listing the 14 pathologies along with a "No Finding" option. Prompt design was guided by established best practices in medical AI prompting, incorporating structured response constraints and terminology standardization to minimize ambiguity and improve classification reliability [ ]. The standardized prompt proceeded as follows:

This is a Chest X-ray image from a patient. Based on the image, does the patient have:

A) Aortic enlargement

B) Atelectasis

C) Calcification

D) Cardiomegaly

E) Consolidation

F) Interstitial lung disease (ILD)

G) Infiltration

H) Lung Opacity

I) Nodule/Mass

J) Other lesion

K) Pleural effusion

L) Pleural thickening

M) Pneumothorax

O) Pulmonary fibrosis

N) No Finding

The model was prompted to identify the presence of specific pathologies in a given CXR image and respond using only the corresponding letter codes assigned to each pathology. For example, if the model detected aortic enlargement, infiltration, and lung opacity, it would return the output "A, G, H"—without any additional explanation or commentary.

To ensure consistency, we automated the process by running each image through a standardized query loop. This involved systematically feeding images to the model one at a time, collecting its responses in a structured format, and storing the results for later analysis. By using an automated approach, we eliminated variations in how prompts were delivered and ensured that each image was evaluated under identical conditions.

Following model output collection, we systematically assessed classification performance using key diagnostic metrics. Accuracy was calculated as the proportion of correctly classified cases relative to the total number of cases. Precision, or positive predictive value, measured the proportion of true positive cases among all positive predictions, while negative predictive value reflected the probability of a true negative result given a negative classification by the model. The F1-score, computed as the harmonic mean of precision and recall, provided a balanced evaluation of classification performance. Additionally, support values were calculated to estimate the prevalence of each pathology in the dataset. To further analyze classification trends and misclassifications, we generated a multi-class confusion matrix (figure 1), allowing for the identification of error distributions, false positives, and potential biases in the model's diagnostic output. This framework enabled us to comprehensively evaluate the strengths and limitations of ChatGPT-4o mini in multi-class pathology classification and determine its feasibility for AI-assisted radiological interpretation.

Results and Discussion:

In applying the ChatGPT-4o mini API to CXR pathology classification, we identified several challenges in differentiating between multiple classes, particularly among underrepresented disease categories. Notably, pneumothorax (PTX), interstitial lung disease (ILD), and aortic enlargement had among the lowest support values in the dataset (e.g., PTX had only 19 cases), leading to complete misclassification with F1-scores of 0.00. The model's overall accuracy of 0.05 suggests that classifying the subset of 14 pathologies posed a considerable challenge, while the macro-averaged precision of 0.28, recall of 0.17, and F1-score of 0.14 indicate only slight improvements over random classification.

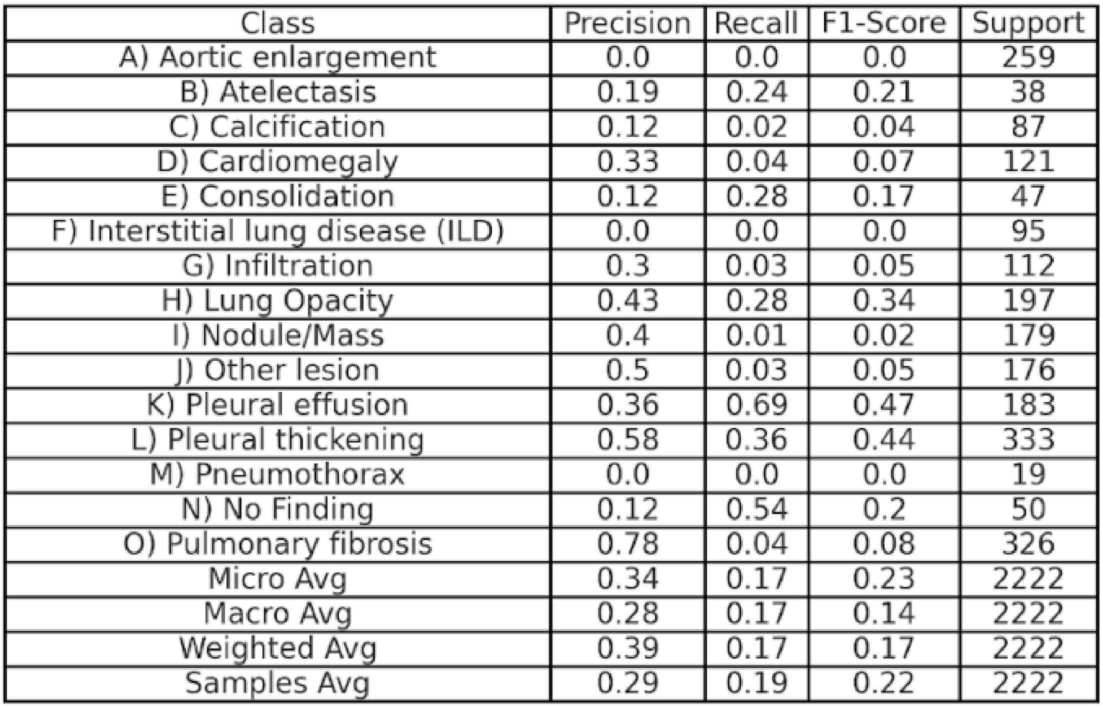
While dataset imbalance played a major role in poor classification performance for underrepresented pathologies, classification failures extended beyond frequently occurring conditions. Even pathologies with relatively higher support values, such as pleural thickening and pleural effusion, exhibited suboptimal performance due to overlapping radiographic features, model uncertainty, and potential biases in label representation. These findings indicate that performance concerns are not limited to dataset prevalence but also to the model's ability to learn distinguishing characteristics among pathologies with shared imaging manifestations.

Table 1 summarizes ChatGPT-4o mini's classification performance in detecting 14 CXR pathologies, highlighting substantial deficiencies in multi-class classification. Precision and recall values varied significantly across different pathology classes. Pulmonary fibrosis (Class O) exhibited the highest precision (0.78), indicating that the model was highly selective in its positive classifications, rarely misidentifying other conditions as pulmonary fibrosis. This class comprised X cases, representing Y% of the examined dataset. Given its relatively low prevalence, the model's precision may be influenced by fewer instances for classification, potentially explaining its low recall (0.04) despite high precision. However, its recall of 0.04 demonstrates a severe limitation in identifying true positive cases, suggesting a preference for conservative classification at the cost of sensitivity. This trade-off is problematic in clinical settings as it would lead to missed diagnoses of pulmonary fibrosis and delays in necessary medical interventions.

Class	Precision	Recall	F1-Score	Support (Cases per pathology)	Dataset Composition
A) Aortic enlargement	0.00	0.00	0.00	259	11.66%
B) Atelectasis	0.19	0.24	0.21	38	1.71%
C) Calcification	0.12	0.02	0.04	87	3.92%
D) Cardiomegaly	0.33	0.04	0.07	121	5.45%
E) Consolidation	0.12	0.28	0.17	47	2.12%
F) Interstitial Lung Disease (ILD)	0.00	0.00	0.00	95	4.28%
G) Infiltration	0.30	0.03	0.05	112	5.04%
H) Lung Opacity	0.43	0.28	0.34	197	8.87%
I) Nodule/ Mass	0.40	0.01	0.02	179	8.06%
J) Other Lesion	0.50	0.03	0.05	176	7.92%
K) Pleural Effusion (PE)	0.36	0.69	0.47	183	8.24%
L) Pleural	0.58	0.36	0.44	333	14.99%

Thickening					
M) Pneumothorax (PTX)	0.00	0.00	0.00	19	0.86%
N) Pulmonary Fibrosis	0.78	0.04	0.08	326	14.67%
O) No Finding	0.12	0.54	0.20	50	2.25%
Micro Avg	0.34	0.17	0.23	2222	-
Macro Avg	0.28	0.17	0.14	2222	-
Weighted Avg	0.39	0.17	0.17	2222	-
Samples Avg	0.29	0.19	0.22	2222	-

Overall Accuracy	0.05
Precision (Macro)	0.28
Recall (Macro)	0.17
F1 Score (Macro)	0.14



Class	Precision	Recall	F1-Score	Support
A) Aortic enlargement	0.0	0.0	0.0	259
B) Atelectasis	0.19	0.24	0.21	38
C) Calcification	0.12	0.02	0.04	87
D) Cardiomegaly	0.33	0.04	0.07	121
E) Consolidation	0.12	0.28	0.17	47
F) Interstitial lung disease (ILD)	0.0	0.0	0.0	95
G) Infiltration	0.3	0.03	0.05	112
H) Lung Opacity	0.43	0.28	0.34	197
I) Nodule/Mass	0.4	0.01	0.02	179
J) Other lesion	0.5	0.03	0.05	176
K) Pleural effusion	0.36	0.69	0.47	183
L) Pleural thickening	0.58	0.36	0.44	333
M) Pneumothorax	0.0	0.0	0.0	19
N) No Finding	0.12	0.54	0.2	50
O) Pulmonary fibrosis	0.78	0.04	0.08	326
Micro Avg	0.34	0.17	0.23	2222
Macro Avg	0.28	0.17	0.14	2222
Weighted Avg	0.39	0.17	0.17	2222
Samples Avg	0.29	0.19	0.22	2222

Overall Accuracy: 0.05

Precision (Macro): 0.28

Recall (Macro): 0.17

F1 Score (Macro): 0.14

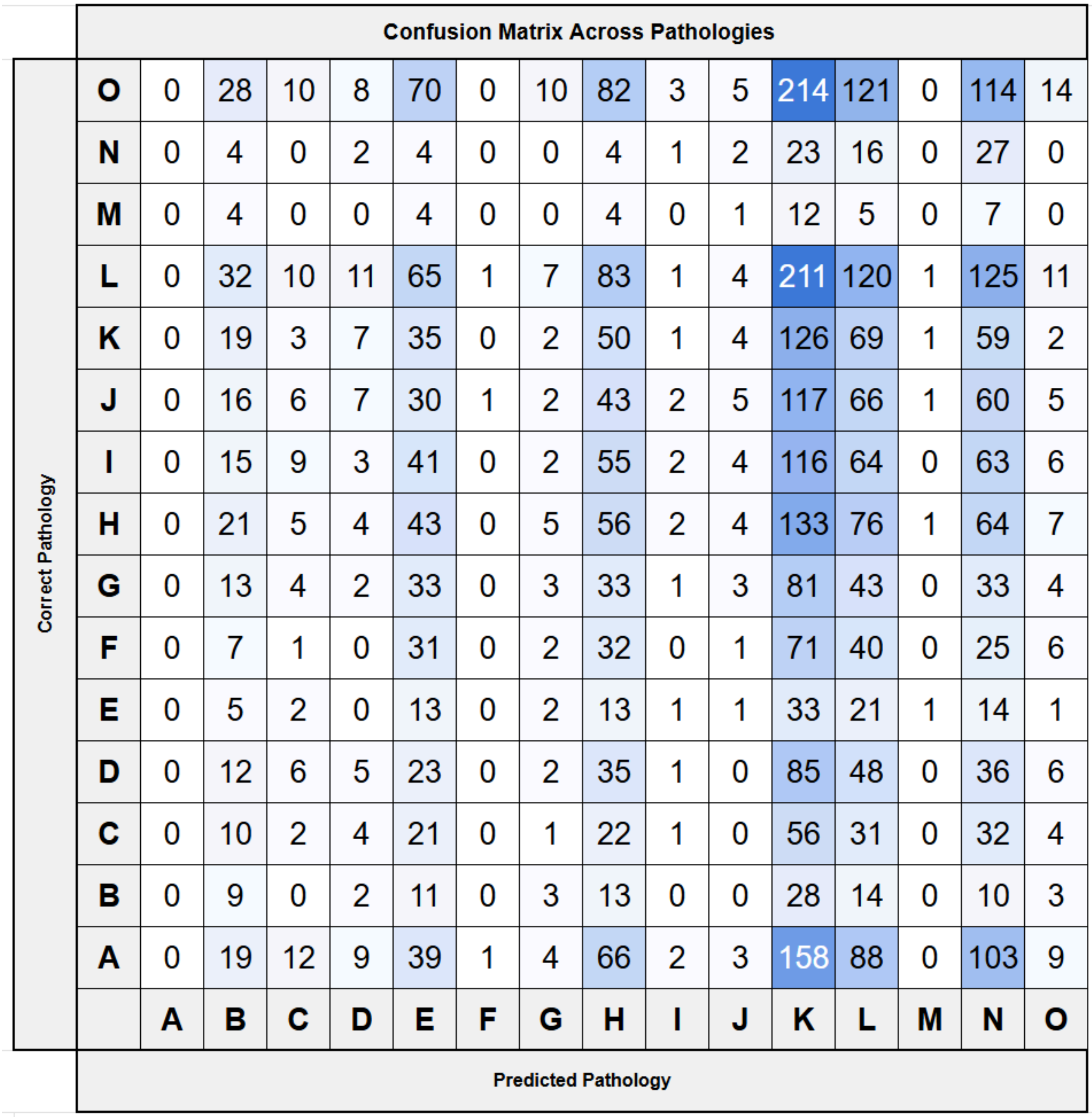
Table 1: Classification Performance Metrics for ChatGPT-4o Mini in Multi-Class CXR Pathology Classification. Table 1: Classification Performance Metrics for ChatGPT-4o Mini in Multi-Class CXR Pathology Classification. A summary of the classification performance of ChatGPT-4o mini in detecting 14 distinct CXR pathologies reveals significant limitations in multi-class classification. While pulmonary fibrosis (Class N, 326 cases, 14.67% of examined dataset) achieved the highest precision (0.78), its low recall (0.04) indicates poor sensitivity. Several conditions, including aortic enlargement (Class A, 259 cases, 11.66% of examined subset), ILD (Class F, 95 cases, 4.28% of examined subset), and PTX (Class M, 19 cases, 0.86% of examined subset), had precision, recall, and F1-scores of 0.00, highlighting a complete failure to classify these conditions correctly.

Pathologies with higher support values, such as pleural thickening (Class K, 183 cases, 8.24% of examined subset) and pleural effusion (Class L, 333 cases, 14.99% of examined subset), demonstrated improved F1-scores, reinforcing the impact of dataset representation on model performance. In contrast, pathologies with low sample sizes (n < X, defined as pathologies comprising less than Y% of the examined dataset), including PTX (Class M, 19 cases, 0.86% of examined subset), ILD (Class F, 95 cases, 4.28% of examined subset), and aortic enlargement (Class A, 259 cases, 11.66% of examined subset), suffered from extreme misclassification rates, further underscoring the model's difficulty in distinguishing pathologies with lower representation. The overall accuracy (0.05) and macro-averaged F1-score (0.14) reflect the model's difficulty in distinguishing multiple pathologies, emphasizing the need for improved dataset balancing and alternative AI architectures to enhance diagnostic reliability.

The generated multi-class confusion matrix (Figure 1) further confirms these shortcomings, with 513 misclassifications for aortic enlargement and no correctly identified cases for aortic dilation (Class A), ILD (Class F), and PTX (Class M). While PTX is typically considered one of the more straightforward pathologies to identify on CXR due to hallmark radiographic signs (e.g., visceral pleural line without lung markings beyond, deep sulcus sign in supine patients), the model’s complete failure suggests deficiencies in its ability to recognize these defining features. This may be attributed to dataset imbalances, inconsistent labeling, or the model’s reliance on text-based inference rather than image-based pattern recognition.

In contrast, both aortic dilation and ILD present greater diagnostic challenges due to their variable radiographic presentations and frequent comorbidities. Aortic dilation often occurs alongside conditions such as cardiomegaly, pleural effusions, or vascular tortuosity, which can obscure clear detection. ILD, on the other hand, is characterized by subtle interstitial changes, including reticular opacities, honeycombing, and ground-glass opacities—patterns that overlap with other pulmonary conditions like pulmonary edema or infections. The model’s inability to correctly classify these conditions suggests that it lacks learned feature representations to differentiate them from visually similar pathologies. Given these findings, further refinement of training methodologies and improved dataset balancing are necessary to enhance model performance, particularly for conditions where radiographic features are subtle or commonly seen in comorbid presentations [ ].

3



Confusion Matrix Across Pathologies

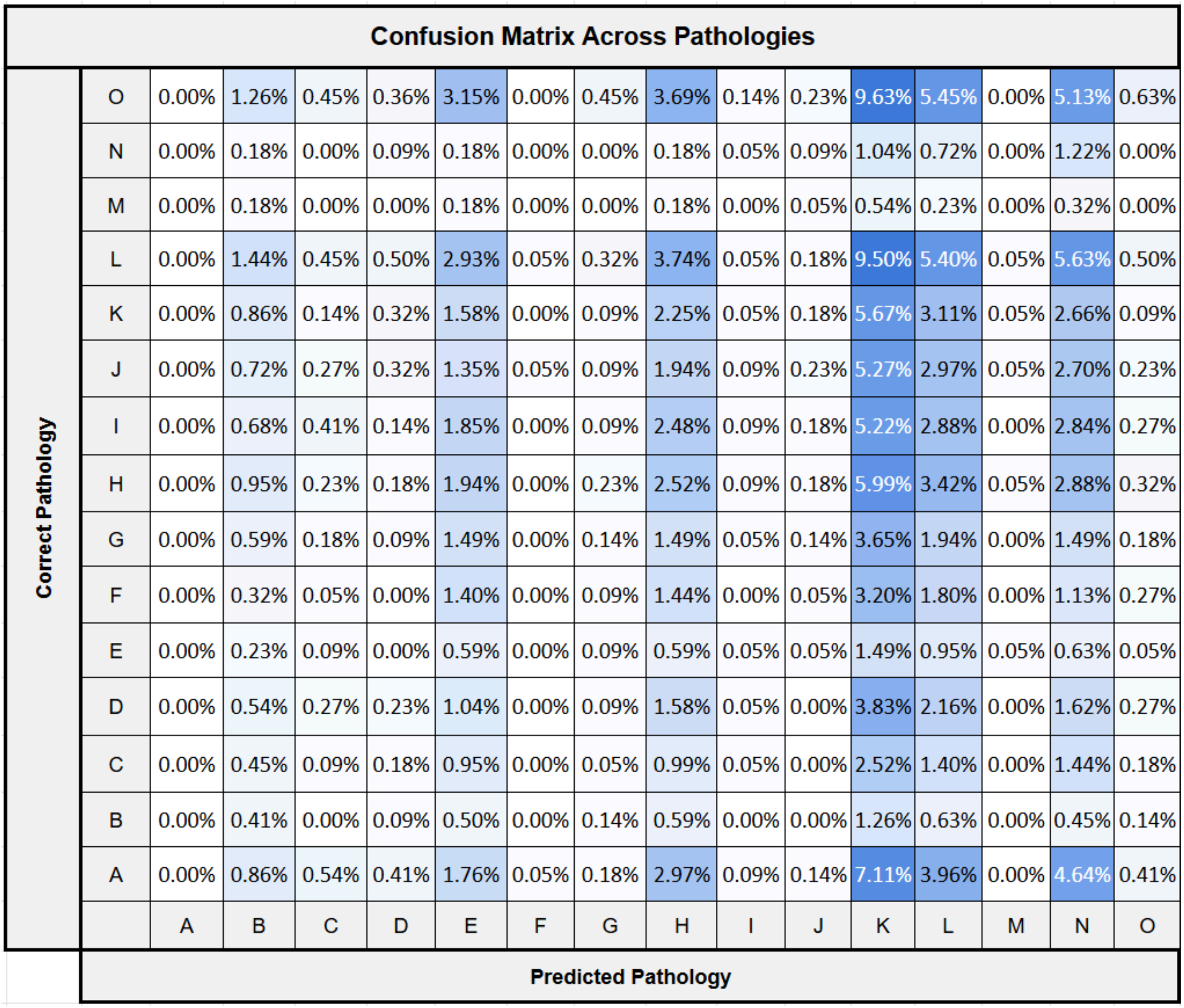
Correct Pathology

O	0	28	10	8	70	0	10	82	3	5	214	121	0	114	14
N	0	4	0	2	4	0	0	4	1	2	23	16	0	27	0
M	0	4	0	0	4	0	0	4	0	1	12	5	0	7	0
L	0	32	10	11	65	1	7	83	1	4	211	120	1	125	11
K	0	19	3	7	35	0	2	50	1	4	126	69	1	59	2
J	0	16	6	7	30	1	2	43	2	5	117	66	1	60	5
I	0	15	9	3	41	0	2	55	2	4	116	64	0	63	6
H	0	21	5	4	43	0	5	56	2	4	133	76	1	64	7
G	0	13	4	2	33	0	3	33	1	3	81	43	0	33	4
F	0	7	1	0	31	0	2	32	0	1	71	40	0	25	6
E	0	5	2	0	13	0	2	13	1	1	33	21	1	14	1
D	0	12	6	5	23	0	2	35	1	0	85	48	0	36	6
C	0	10	2	4	21	0	1	22	1	0	56	31	0	32	4
B	0	9	0	2	11	0	3	13	0	0	28	14	0	10	3
A	0	19	12	9	39	1	4	66	2	3	158	88	0	103	9
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**

Predicted Pathology

Figure 1A: Confusion Matrix Illustrating ChatGPT-4o Mini's Performance in Multi-Class CXR Pathology Classification.

This matrix (**Figure 1A**) visualizes ChatGPT-4o mini's classification performance across the examined subset of 14 CXR pathologies and a "No Finding" (NF) category, highlighting severe misclassification trends. Aortic enlargement (A), PTX (M), and ILD (F) were almost entirely misclassified, with no true positive predictions. Pathologies with higher support values, such as PE (K) and pleural thickening (L), demonstrated comparatively better classification performance. The absence of true positives for PTX is unexpected given its distinct radiographic features. This raises the question of whether the model failed to assign a label in some cases or defaulted to NF when uncertain. The dataset included both pathology-positive and NF cases, yet the model's misclassification patterns suggest potential bias toward NF for challenging pathologies. Further analysis of NF predictions is warranted to determine if ChatGPT-4o mini exhibited label omission tendencies rather than direct misclassification.



Confusion Matrix Across Pathologies

Correct Pathology

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
O	0.00%	1.26%	0.45%	0.36%	3.15%	0.00%	0.45%	3.69%	0.14%	0.23%	9.63%	5.45%	0.00%	5.13%	0.63%
N	0.00%	0.18%	0.00%	0.09%	0.18%	0.00%	0.00%	0.18%	0.05%	0.09%	1.04%	0.72%	0.00%	1.22%	0.00%
M	0.00%	0.18%	0.00%	0.00%	0.18%	0.00%	0.00%	0.18%	0.00%	0.05%	0.54%	0.23%	0.00%	0.32%	0.00%
L	0.00%	1.44%	0.45%	0.50%	2.93%	0.05%	0.32%	3.74%	0.05%	0.18%	9.50%	5.40%	0.05%	5.63%	0.50%
K	0.00%	0.86%	0.14%	0.32%	1.58%	0.00%	0.09%	2.25%	0.05%	0.18%	5.67%	3.11%	0.05%	2.66%	0.09%
J	0.00%	0.72%	0.27%	0.32%	1.35%	0.05%	0.09%	1.94%	0.09%	0.23%	5.27%	2.97%	0.05%	2.70%	0.23%
I	0.00%	0.68%	0.41%	0.14%	1.85%	0.00%	0.09%	2.48%	0.09%	0.18%	5.22%	2.88%	0.00%	2.84%	0.27%
H	0.00%	0.95%	0.23%	0.18%	1.94%	0.00%	0.23%	2.52%	0.09%	0.18%	5.99%	3.42%	0.05%	2.88%	0.32%
G	0.00%	0.59%	0.18%	0.09%	1.49%	0.00%	0.14%	1.49%	0.05%	0.14%	3.65%	1.94%	0.00%	1.49%	0.18%
F	0.00%	0.32%	0.05%	0.00%	1.40%	0.00%	0.09%	1.44%	0.00%	0.05%	3.20%	1.80%	0.00%	1.13%	0.27%
E	0.00%	0.23%	0.09%	0.00%	0.59%	0.00%	0.09%	0.59%	0.05%	0.05%	1.49%	0.95%	0.05%	0.63%	0.05%
D	0.00%	0.54%	0.27%	0.23%	1.04%	0.00%	0.09%	1.58%	0.05%	0.00%	3.83%	2.16%	0.00%	1.62%	0.27%
C	0.00%	0.45%	0.09%	0.18%	0.95%	0.00%	0.05%	0.99%	0.05%	0.00%	2.52%	1.40%	0.00%	1.44%	0.18%
B	0.00%	0.41%	0.00%	0.09%	0.50%	0.00%	0.14%	0.59%	0.00%	0.00%	1.26%	0.63%	0.00%	0.45%	0.14%
A	0.00%	0.86%	0.54%	0.41%	1.76%	0.05%	0.18%	2.97%	0.09%	0.14%	7.11%	3.96%	0.00%	4.64%	0.41%

Predicted Pathology

Figure 1B: Confusion Matrix in Figure 1A as percentages, given as cell value, x, as a percentage of total cases, X = 2222.

To improve reliability, a simplified confusion matrix including only pathologies with >10% representation would provide a clearer assessment of classification trends. These findings are consistent with Table 1 and emphasize the need for dataset balancing or weighted loss functions to mitigate bias toward more prevalent pathologies, reinforcing concerns about the model’s generalizability and the necessity for further refinements before clinical application.

These errors appear to be influenced by several key factors, the most significant being dataset imbalance, which directly impacted model performance. ChatGPT-4o mini performed better on pathologies that were more prevalent in the dataset (Pleural Effusion and Pleural Thickening) and struggled with those that were underrepresented (Consolidation and Pneumothorax). For instance, PE (Class K, 183 cases, 8.24% of examined subset) and Pleural Thickening (Class L, 333 cases, 14.99% of examined subset) demonstrated higher recall and F1-scores, suggesting that the sample distribution in the training data played a critical role in model effectiveness. In contrast, PTX (Class M, 19 cases, 0.86%) was frequently misclassified, which is unexpected given its well-documented prevalence and characteristic radiographic findings. This misclassification raises concerns about the model's ability to recognize PTX despite its relative clinical frequency. PTX is typically straightforward to identify on CXR due to its distinct radiographic presentation (e.g., absence of lung markings, visible pleural line), yet the model failed to consistently detect it.

	Representation in Dataset, %	Representation in Dataset, Ranking where most prevalent = 1	Representation in the General Population	Reference, Comments
A) Aortic enlargement	11.66%	3	2.7%	(13) Elderly Population
B) Atelectasis	1.71%	14	No relevant data	

C) Calcification	3.92%	11	No relevant data	
D) Cardiomegaly	5.45%	8	23-35.6%	(14) Elderly Population (15)
E) Consolidation	2.12%	13	15%	(16) Under 5 Population
F) Interstitial Lung Disease (ILD)	4.28%	10	0.0086%	(17) Disease Prevalence, not prevalence of CXR findings of ILD
G) Infiltration	5.04%	9	2.15-3.23%	(18) Mean age of patient 13 months
H) Lung Opacity	8.87%	4	5.3%	(19)
I) Nodule/ Mass	8.06%	6	0.09%–0.20%	(20)
J) Other Lesion	7.92%	7	No relevant Data	
K) Pleural Effusion (PE)	8.24%	5	0.34%	(21) Incidence in General Population
L) Pleural Thickening	14.99%	1	35.2%	(22) General Health Check CXR
M) Pneumothorax (PTX)	0.86%	15	0.017%	(23) Incidence in General Population
N) Pulmonary Fibrosis	14.67%	2	No relevant Data	
O) No Finding	2.25%	12	No relevant Data	

Table 2: A comparison of the relative contribution of each pathology from the Multi-Class CXR Classification compared to their approximate incidence/ prevalence in the relevant literature.

One possible explanation is the limited representation of PTX in our examined subset—comprising only 19 cases—compared to its broader prevalence in the full VinDr-CXR dataset. If the subset underrepresents PTX relative to its true distribution in clinical practice, the model's performance could be skewed due to inadequate exposure during evaluation. However, if PTX misclassification persists despite sufficient representation in the full dataset, this would indicate a fundamental issue in the model's decision-making rather than a sampling artifact. This distinction is critical in assessing whether the misclassification stems from dataset composition or inherent limitations in LLMs for radiographic interpretation.

Beyond dataset imbalance, these results underscore the inherent limitations of using LLMs for direct CXR interpretation. Unlike CNNs, which extract pixel-level features, ChatGPT-4o mini relies solely on text-based reasoning, making it ill-equipped to analyze radiographic patterns. This shortcoming was particularly evident for certain pathologies, where misclassification extended beyond dataset prevalence issues to fundamental failures in feature recognition. Notably, the absence of true positives for multiple classes—including PTX—suggests that key radiographic features were potentially unrecognized entirely.

While the inclusion of lower-representation classes (B, M, N) may have contributed to misclassification trends, it also serves as a valuable insight into LLM limitations when handling underrepresented disease categories. This highlights the necessity of evaluating LLMs across diverse pathology distributions rather than exclusively focusing on high-prevalence conditions. Future work should refine model training methodologies, incorporating more balanced datasets or tailored learning strategies to enhance generalizability across all pathology classes.

Limitations and Future Directions

In this study, we explored the limitations of utilizing LLMs such as ChatGPT-4o mini for CXR pathology classification. A key limitation of ChatGPT-4o mini is its reliance on text-based reasoning rather than direct image analysis, which constrains its ability to capture spatial patterns critical for radiographic interpretation. This challenge is particularly evident in conditions with subtle, overlapping imaging features (e.g., ILD, aortic dilation) and in cases where the model appears to default to "No Finding" rather than assigning uncertain classifications. Given these constraints, improving LLM-based approaches for CXR interpretation will require integration with models capable of pixel-level feature extraction, such as CNNs, to mitigate these shortcomings.

Biases in the model’s training dataset, including inconsistencies in class distributions and labeling, may have contributed to misclassification errors. Prior studies have demonstrated that AI-based radiology models often exhibit biases favoring more frequently co-occurring pathologies or those with higher radiographic contrast, leading to challenges in differentiating conditions with more subtle imaging features [1,2]. While PTX is a well-documented, relatively common pathology with distinct radiographic signs, its misclassification in this study suggests deficiencies in label reliability, model decision-making, or reliance on text-based inference rather than direct spatial analysis.

To enhance classification accuracy, dataset curation must ensure that pathologies are proportionally represented while maintaining diagnostic diversity. Additionally, model architecture refinements—such as integrating deep learning models optimized for image-based pattern recognition—could improve feature extraction and differentiation among overlapping pathologies. Future work should explore multimodal AI frameworks that combine text-based reasoning with pixel-level feature extraction to enhance diagnostic performance across a wider range of conditions.

Additionally, classification techniques need further refinement, as discrepancies observed in Figure 1 indicate the model struggles to differentiate between pathology classes. One potential approach is implementing class-specific loss functions, which adjust how the model learns based on the difficulty of identifying each pathology. For example, if a rare condition like ILD is often misclassified, the model can be programmed to 'penalize' incorrect predictions for ILD more heavily than for a more common condition like pleural effusion. This adjustment forces the model to focus more on learning the distinguishing features of underrepresented conditions. Similarly, advanced contrastive learning techniques—which train the model to better recognize subtle differences between similar-looking pathologies—could improve overall classification accuracy. Refining these methods could enhance sensitivity to infrequent but clinically significant conditions, ultimately increasing generalizability and making the model more reliable for real-world use.

Conclusions

ChatGPT-4o mini exhibits several limitations in CXR pathology classification, particularly in a multi-class setting where class imbalance poses a significant challenge. While the model demonstrated high precision in identifying pulmonary fibrosis, its failure to detect critical conditions such as PTX, ILD, and aortic enlargement suggests that it is not yet suitable as a standalone diagnostic tool. These findings align with broader challenges in AI-based radiology and emphasize the necessity of thorough validation and dataset balancing before such models can be integrated into clinical practice. Based on these results, radiographic assessment should unsurprisingly remain the gold standard for clinical decision-making, with AI models like ChatGPT-4o mini serving only as supplementary tools until further refinement and validation are achieved.

References:

1. Jones CM, Buchlak QD, Oakden-Rayner L, Milne M, Seah J, Esmaili N, Hachey B. Chest radiographs and machine learning - Past, present and future. J Med Imaging Radiat Oncol. 2021 Aug;65(5):538-544. doi: 10.1111/1754-9485.13274. Epub 2021 Jun 25. PMID: 34169648; PMCID: PMC8453538.
2. Najjar R. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. Diagnostics (Basel). 2023 Aug 25;13(17):2760. doi: 10.3390/diagnostics13172760. PMID: 37685300; PMCID: PMC10487271.
3. Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8
4. Babar M, Qureshi B, Koubaa A. Investigating the impact of data heterogeneity on the performance of federated learning algorithm using medical imaging. PLoS One. 2024 May 15;19(5):e0302539. doi: 10.1371/journal.pone.0302539. PMID: 38748657; PMCID: PMC11095741.
5. Chamseddine E, Mansouri N, Soui M, Abed M. Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss. Appl Soft Comput. 2022 Nov;129:109588. doi: 10.1016/j.asoc.2022.109588. Epub 2022 Aug 29. PMID: 36061418; PMCID: PMC9422401.

6. Zhang, Z., Zhang, X., Ichiji, K. et al. How intra-source imbalanced datasets impact the performance of deep learning for COVID-19 diagnosis using chest X-ray images. Sci Rep 13, 19049 (2023). https://doi.org/10.1038/s41598-023-45368-w
7. Duhaime, E. P., Jin, M., Moulton, T., Weber, J., Kurtansky, N. R., & Halpern, A. (2023). Nonexpert crowds outperform expert individuals in diagnostic accuracy on a skin lesion diagnosis task. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1–4). IEEE. https://doi.org/10.1109/ISBI53787.2023.10230646

8. Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Nguyen, D. T., Nguyen, N. T., Nguyen, V. V., Dao, L. H., Vu, N. M., Tran, N. K., Nguyen, H. Q., Tran, T. B., Phi, C. D., Do, C. D., Nguyen, H. T., Nguyen, P. H., Nguyen, A. V., . . .Vu, V. (2022). VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. Scientific Data, 9(1), 429
9. Wang, C. M., Elazab, A., Wu, J. H., & Hu, Q. M. (2017). Lung nodule classification using deep feature fusion in chest radiography. Computerized Medical Imaging and Graphics, 57, 10-18
10. Tiu, E., Talius, E., Patel, P., Curtis, C., Ward, C., Ades, S., Tran, T., Ngo, P., Gillies, R., Patel, T., Goldgof, D., Hall, L., Drukker, K., Giger, M., Wolfson, S., Freymann, J., Kirby, J., Jaffe, C., Maidment, A., . . .Summers, R. M. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering, 6(12), 1399-1406
11. Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W. J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. Engineering Applications of Artificial Intelligence, 92, 103678
12. Chen, Y., Wan, Y., & Pan, F. (2023). Enhancing Multi-disease Diagnosis of Chest X-rays with Advanced Deep-learning Networks in Real-world Data. Journal of Digital Imaging, 36(4), 1332-1347.