Article

Not peer-reviewed version

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

Swapna Vaja [*] , Shivam Patel , Amogha Chetla , Diana Chen , Kunal Sukhija

Posted Date: 6 December 2024

doi: 10.20944/preprints202412.0549.v1

Keywords: Artificial intelligence; radiology; pathology

Article

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

Swapna Vaja [1,*], Shivam Patel [2], Amogha Chetla [3], Diana Chen [4] and Kunal Sukhija [5]

[1] Rush Medical College, USA
[2] University of Virginia, USA
[3] Independent Researcher, USA
[4] Independent Researcher, USA
[5] Independent Researcher, USA
* Correspondence: swapna.vaja@gmail.com

Abstract: Radiographic assessment of chest X-rays (CXR) is the current standard for diagnostic imaging, but advancements in artificial intelligence (AI) have opened new possibilities for automated pathology detection. This study evaluates the performance of OpenAI's ChatGPT-4o mini in correctly identifying 14 distinct chest X-ray pathologies or confirming the absence of pathology using the VinDr-CXR dataset - specifically through the Kaggle data subset. After the GPT model was queried with a standardized prompt, key performance metrics like accuracy, precision, recall, and F1 scores were calculated, and a multi-classification confusion matrix were analyzed to assess performance. Results revealed significant limitations in ChatGPT-4o mini's diagnostic capabilities, with an overall accuracy of 0.05 and macro-averaged precision, recall, and F1 scores of 0.28, 0.17, and 0.14, respectively. Moreover, several pathologies, including aortic enlargement, interstitial lung disease, and pneumothorax, were entirely misclassified. Performance variability across classes appeared associated with dataset imbalances, as classes with higher support values generally showed more favorable outcomes. These findings highlight the significant challenges faced by the current ChatGPT-4o mini model in multi-class diagnostic classification, underscoring a need for improved model training before successful integration into practical clinical scenarios can be undertaken.

Keywords: artificial intelligence; radiology; pathology

Introduction

Radiographic assessment of chest X-ray scans is presently considered the first line for imaging interpretation; however, with recent advancements in AI-driven diagnostics [1], our study aims to elucidate the capabilities of such technologies, such as OpenAI's ChatGPT-o mini, in detecting chest X-ray pathologies. Large language models (LLMs) are AI-based technologies trained on large amounts of data to simulate intelligence - offering demonstrable clinical use cases [2]. With the necessity of utilizing high quality datasets [3], as well as their increasing availability, the ability to rigorously evaluate ChatGPT's clinical reliability is higher than ever. Our study aims to evaluate that reliability by leveraging the VinDr-CXR dataset formulated by Nguyen et al. [1] VinDr-CXR is comprised of 18,000 chest X-ray images annotated by experienced radiologists for a broad range of 27 pathologies including, but not limited to: pneumonia, tuberculosis, and lung tumor.

Methods

Our study evaluated the performance of OpenAI's API for ChatGPT-4o mini in classifying 14 clinically significant pathologies or correctly identifying the absence of pathology from chest X-ray images. The VinDR-CXR dataset contains 18,000 annotated chest X-ray images spanning 27 pathologies, but only pathologies and images from the Kaggle train subset - which contained 14 pathologies and 15,000 images - were included. The dataset was restructured so that each image had an annotated row of present pathologies with the number of occurrences of each pathology. Out of the 15,000 images that were a part of the Kaggle subset, 568 images had a dominant pathology present of at least 3 occurrences higher than the next dominant pathology. The scans included in the Kaggle subset were used without any sort of modification or annotation on our end. Images were processed through a recursive Python loop, which queried the API with the prompt "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. Give a list of letters. Give nothing else but the letters. For example, if you believe it's aortic enlargement, Infiltration, and Lung Opacity, you would respond with 'A, G, H', and nothing else."

Key performance metrics, including accuracy, precision, recall, F1 score, and support values, were computed. Additionally, a combined multiclass confusion matrix for all of the used pathologies was generated to visualize classification performance.

Results and Discussion

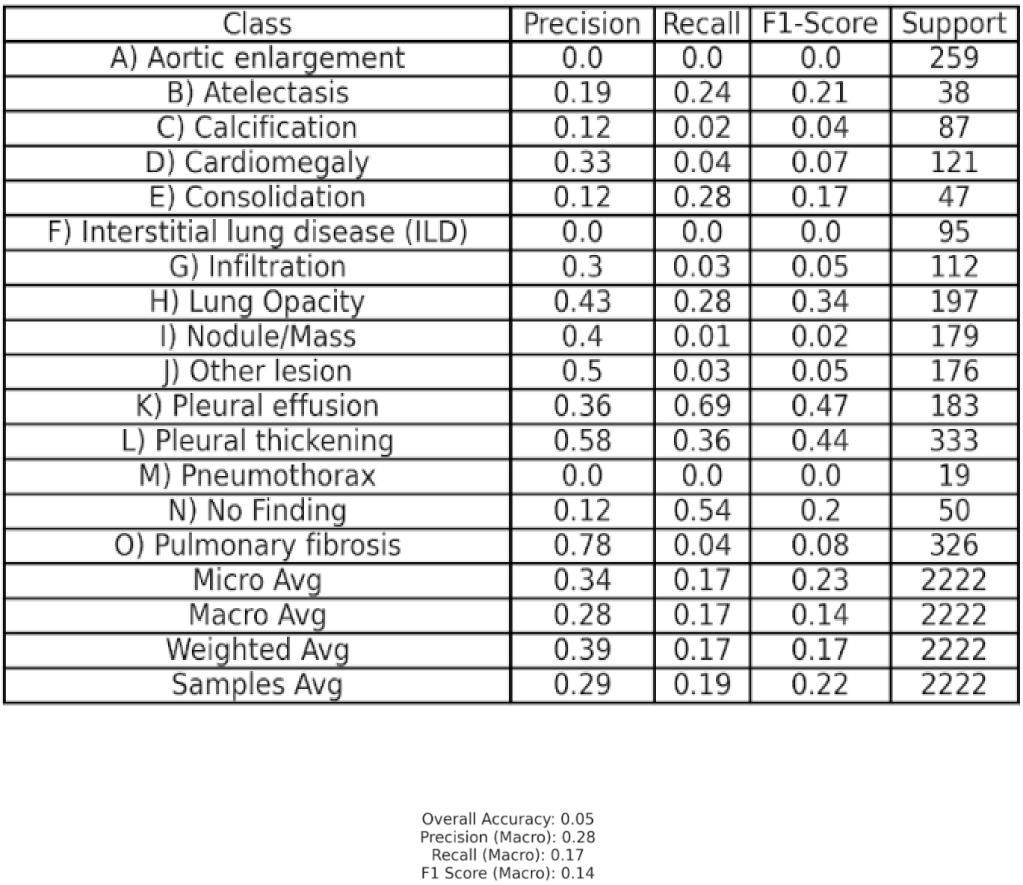
Our results demonstrate that GPT-4o mini presents with poor performance in correctly classifying the presence of different pathologies. Based on an overall accuracy score of 0.05, it can be inferred that GPT-4o mini experiences significant difficulty in correctly differentiating between the 14 classes of chest X-ray pathologies presented to it. Macro-averages of 0.28, 0.17, and 0.14 across metrics of precision, recall, and F1 score are marginal improvements over the recorded accuracy but are still indicative of poor predictive performance. GPT-4o mini demonstrated significant variance in performance between classes of diseases as well - notably, Class O (pulmonary fibrosis) presented the highest measure of precision at 0.78, suggesting GPT-4o mini made few false-positive predictions. However, Class O's recall of 0.04 indicated that GPT-4o mini failed to identify the vast majority of true instances of pulmonary fibrosis presented to it.

Classes A, F, and M - representing aortic enlargement, Interstitial Lung Disease (ILD), and pneumothorax respectively - all had F1 scores of 0.00, indicating total failure to identify any presented images correctly. These findings are further corroborated by results such as 513 misclassifications of aortic enlargement (Class A) as shown by the multi-class confusion matrix. With a total lack of true positive and false positive responses, there is demonstrable evidence that GPT-4o mini does not offer significant discriminatory pattern recognition for the identification of aortic enlargement on CXR, a pattern generalizable to ILD and pneumothorax. Given the cruciality of minimizing false positives and maximizing true positive detection rates [4], we express strong reservations in the current GPT-o mini model's clinical capabilities.

The disparities present in identification performance across classes may be indicative of imbalances within the dataset, as evidenced by trends associated with support values. Classes with high support values (SV) like Class K (pleural thickening with SV=183) and Class L (pleural effusion with SV=333) seemed to generally present with more rigorous metrics of success across measures of precision, recall, and F1 score. Conversely, identification classes like Class M (pneumothorax with SV = 19) may not have enough data to support a rigorous conclusion of GPT-o mini's true capabilities.

Overall, GPT-4o mini's current model demonstrates significant limitations in its ability to make correct identifications across 15 CXR interpretation outcomes, as evidenced by poor statistical measures of virtually all measures of competence (i.e precision, recall, F1 score, and accuracy) both across and within classes of pathologies presented. Our findings corroborate a broader sentiment of artificial intelligence models struggling to accurately interpret multi-classification-based datasets [5]. Our current recommendation is for physicians to rely on radiographic assessment to make clinical judgments when evaluating potential pathologies on CXR.

Table 1. Classification performance metrics for ChatGPT4-o mini's performance in assessing chest X-ray pathologies. Metrics include precision, recall, and F1-score for each class, along with the total class support. Aggregated metrics—micro, macro, weighted, and sample averages—are included to provide an overall assessment.



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

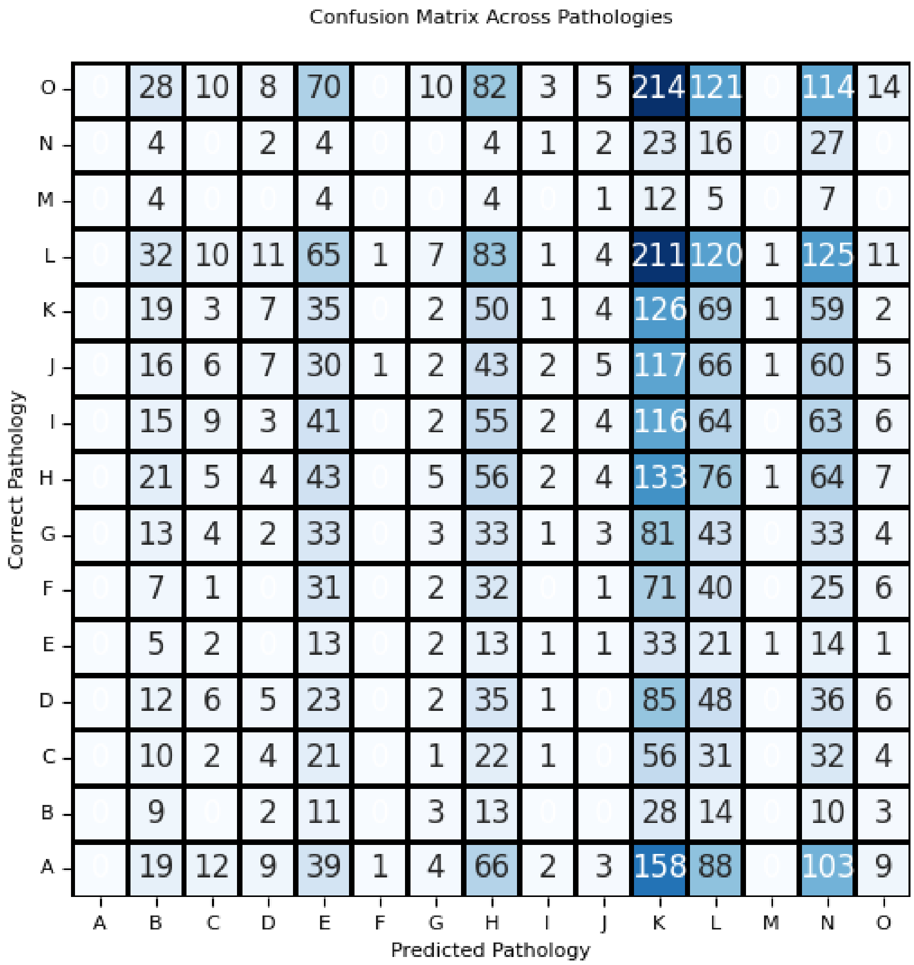


Figure 1. The performance of ChatGPT-4o mini in classifying CXR pathologies has been visualized through the creation of a confusion matrix. Notable lapses in performance can be observed in the nearly total misidentification of pathologies like aortic enlargement (A), pneumothorax (M), and ILD (F).

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.