Evaluating Large Multimodal Models in COVID-19 Pneumonia Detection: A Case Study Using Chest X-Rays

Co-authors: Nitin Chetla, Tamer Hage, Swapna Vaja, Adarsh Mallepally, Harshita Kacham, Yasmeen Abisaab, Shivam Patel, Sai Samayamanthula, Varun Raja, Rahul Reddy, Kunal Sukhija

Introduction:

Patients severely afflicted with coronavirus disease (COVID-19) commonly experience pneumonia, making radiographs an essential component of the diagnostic workup [1]. In fact, the rapid identification and diagnosis of COVID-19 pneumonia from chest X-rays (CXR) is critical for timely clinical management and resource allocation [2]. As such, researchers in recent years have leveraged deep learning (DL)-based solutions using convolutional neural networks (CNNs) for medical imaging analysis, ushering innovative opportunities to transform a widely used diagnostic tool for COVID-19 [3,4]. These DL models can extract complex features that are otherwise not easily visible, aiding radiologists in distinguishing between COVID-19 pneumonia and other viral pneumonias or pulmonary disease states [5]. Prior studies proposing DL models have achieved accuracy scores upwards of 90%, demonstrating their clinical utility in not only detecting COVID-19 pneumonia but also classifying its severity [3].

While DL models boast high diagnostic precision, they present challenges with respect to interpretability and transparency, often lacking clear justifications for their diagnostic recommendations [6]. The advent of generative artificial intelligence (AI) platforms, especially large language models (LLMs) with image interpretation ability such as Gemini 2.0, offers a potential solution by providing greater clarity and feedback for users [6]. Compared to other LLMs like ChatGPT, which typically rely on distinct vision encoders to process images, Gemini 2.0 was designed with a more unified multimodal architecture that natively integrates visual and textual understanding. This difference allows Gemini 2.0 to process images and language inputs in a more cohesive manner, whereas traditional LLMs often treat image analysis as a separate pipeline appended to a primarily text-based model. Naturally, these architectural distinctions have stirred interest for their promising ability to enhance diagnostic accuracy and assist healthcare providers. Given this context, we sought to evaluate the effectiveness of Gemini 2.0 in classifying COVID-19 pneumonia from CXRs.

Methods:

We used 20,000 CXR images from the publicly available **COVIDx CXR-4 dataset** [11], equally divided into two groups: COVID-19 pneumonia positive (n=10,000) and negative (n=10,000). Each CXR image was individually submitted to Gemini 2.0's API using a standardized prompt: "This is a chest x-ray that appeared in the Step 1 Medical School Examination. Please answer this question with a single letter ONLY. For example, if you believe there is Covid-19 Pneumonia, answer 'A' and NOTHING ELSE. Does this chest x-ray display evidence of Covid-19 Pneumonia? A) Covid-19 Pneumonia is present B) Covid-19 Pneumonia is not present." We subsequently recorded Gemini 2.0's responses and analyzed the results to calculate accuracy, precision, recall, F1-score, and support.

Results:

Gemini 2.0 achieved an overall accuracy of 45% in distinguishing COVID-19 pneumonia (**Table 1**). Specifically, the precision was 34% for pneumonia-positive cases and 47% for pneumonia-negative cases. Recall was significantly lower in the pneumonia-positive group (11%) compared to the pneumonia-negative group (79%). Consequently, the F1-score was considerably higher for pneumonia-negative cases (0.59) than pneumonia-positive cases (0.17). Out of 10,000 pneumonia-positive images, Gemini correctly identified only 1,119, while incorrectly classifying 8,881 as negative (**Figure 1**). Conversely, among 10,000 pneumonia-negative images, Gemini correctly identified 7,875 and misclassified 2,125 as positive.

Discussion:

The study demonstrates substantial limitations in Gemini 2.0's capability to accurately detect COVID-19 pneumonia from CXR, with particularly poor sensitivity. Additionally, this work builds off previous work by our group using ChatGPT-4 Turbo (54.1% accuracy) and ChatGPT-4o (61.2% accuracy) on the same dataset, ultimately revealing that Gemini 2.0 performed substantially worse at only 45% accuracy. [10]

The high false-negative rate suggests Gemini 2.0, in its current form, is unreliable for clinical diagnosis or screening purposes. This finding sheds light on Gemini 2.0's lack of generalizability, a key challenge that lies in AI models due to their inability to adapt to variance such as changes in image quality, diverse patient populations, and anatomical differences [7]. Additionally, like ChatGPT-4o, generative AI models can experience difficulty with more nuanced feature extraction specific to medical imaging and, more specifically, pneumonia detection [6]. As such, clinicians ought to be aware of AI's limitations and its need for human oversight.

While LLMs may hold promise for medical imaging and diagnosis, these findings underline the importance of further refining AI models before clinical deployment can be considered. A possible solution involves improving generalizability, which includes efforts to incorporate additional training with specialized medical imaging datasets, more diverse and heterogeneous data, and adversarial training [8]. Another approach may be to combine well-established standard image analysis techniques, such as textural analysis and radiomics, with generative AI models [9]. This hybrid approach has the potential to optimize the strengths of both approaches to generate clinically useful diagnoses without compromising interpretability.

References

[1]https://pmc.ncbi.nlm.nih.gov/articles/PMC7802079/

[2]https://www.mdpi.com/2313-433X/10/10/250

[3]https://pmc.ncbi.nlm.nih.gov/articles/PMC10301527/

[4] https://pmc.ncbi.nlm.nih.gov/articles/PMC9483290/

[5]https://mednexus.org/doi/full/10.1002/cdt3.17

[6]https://pmc.ncbi.nlm.nih.gov/articles/PMC11442562/#CR6

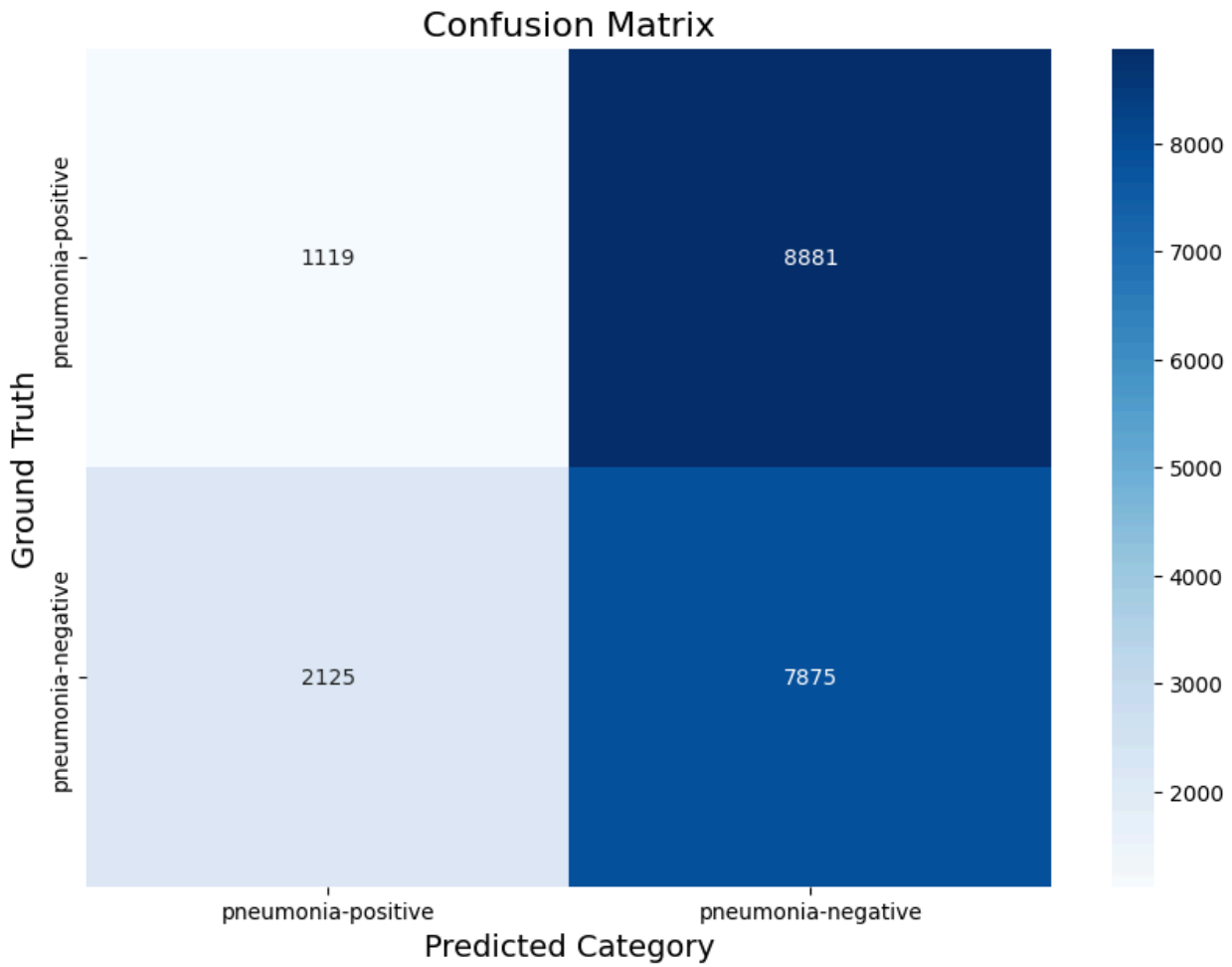
[7]https://www.nature.com/articles/s41746-024-01127-3

[8]https://pmc.ncbi.nlm.nih.gov/articles/PMC8637230/

[9] https://www.mdpi.com/2075-4418/15/3/282

[10] Chetla N, Tandon M, Chang J, Sukhija K, Patel R, Sanchez R. Evaluating ChatGPT's Efficacy in Pediatric Pneumonia Detection From Chest X-Rays: Comparative Analysis of Specialized AI Models. JMIR AI. 2025;4:e67621. Published 2025 Jan 10. doi:10.2196/67621

[11] Wang L, Lin ZQ, Wong A. COVIDx CXR-4: An open-access benchmark dataset for COVID-19 chest X-ray classification. Kaggle. https://www.kaggle.com/datasets/andyczhao/covidx-cxr2. Accessed April 2025.



Metric	Precision	Recall	F1-score	Support
Covid-19 Pneumonia Positive (A)	0.34	0.11	0.17	10000
Covid-19 Pneumonia Negative (B)	0.47	0.79	0.59	10000
Accuracy			0.45	20000
Macro Avg	0.41	0.45	0.38	20000
Weighted Avg	0.41	0.45	0.38	20000