Can Large Language Models Challenge CNNs in Medical Image Analysis?

Shibbir Ahmed[1], Shahnewaz Karim Sakib[2], Anindya Bijoy Das[3]

[1]Texas State University, [2]University of Tennessee Chattanooga, [3]University of Akron

10th ANNUAL HEALTH SCHOLAR SHOWCASE

Summary

- **Comparative study** of CNNs and LLMs (GPT-4o, LLaMA3.2-vision) for medical image classification.
- **Datasets:** X-ray (COVID-19 Radiography), Brain MRI (Tumor), Chest CT Scan (Cancer subtypes).
- **Metrics:** accuracy, F1, confidence score, runtime, energy, CO_2.
- **Key finding**: CNNs offer higher accuracy and efficiency; LLMs can be enhanced data filtering and potential hybrid integration

Problem

- CNNs are **state-of-the-art in imaging** but lack contextual reasoning.
- LLMs promise **multimodal understanding** but remain underexplored in image classification.
- RQ: ***Can LLMs match or complement CNNs in diagnostic imaging?***

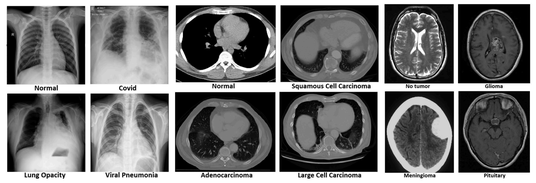


Figure I: Different Medical Imaging Modalities and Pathologies Samples (a) Chest X-ray, (b) Chest CT-Scan, (c) Brain Tumor MRI

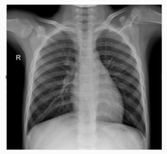
Related Work

- **CNN advances in medical imaging** [Zhou et al., 2021]
- **LLMs in healthcare** [Thirunavukarasu et al., 2023]
- **Hybrid multimodal models** [Radford et al., 2021 – CLIP; Alayrac et al., 2022 – Flamingo]

Approach

- Benchmarked **CNNs** (COVID-Roentgen, Brain Tumor CNN, Chest CT CNN) vs. **LLMs** (GPT-4o, LLaMA3.2-vision).
- Introduced **enhanced data filtering**: label filtering + confidence scores + contextual feature extraction.
- **Metrics:** accuracy, F1, confidence score, execution time, energy, CO_2.
- **Datasets:** COVID-19 Radiography, Bra… Tumor MRI, Chest CT Scan

Please evaluate the chest X-ray image using the following radiological criteria:



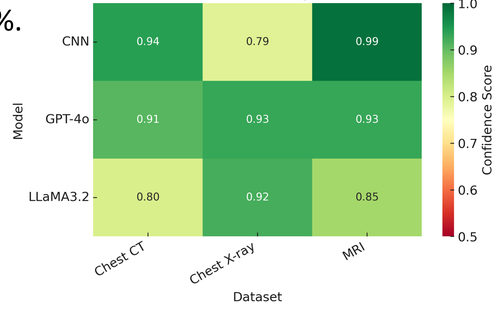
1. Clear Lung Fields
2. Normal Heart and Mediastinum
3. Healthy Diaphragm
4. Intact Bony Structures
5. Empty Pleural Spaces

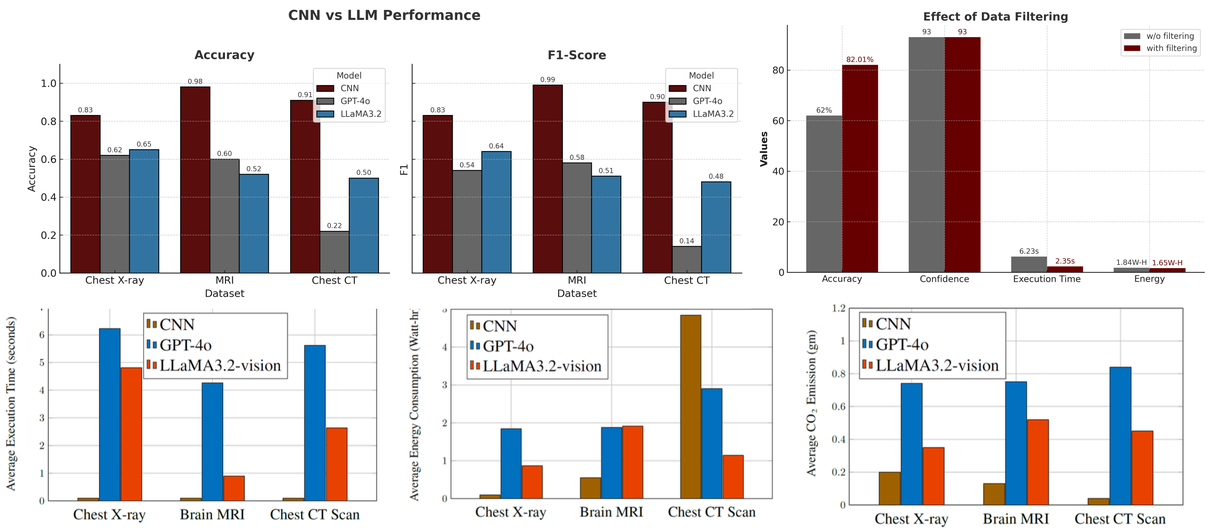
Based on the above criteria, does this image appear normal or COVID-affected?
Please reply in exactly two lines. The first line should contain only either `normal` or `abnormal`. The second line should be in the format `Confidence Score:<number>` where `<number>` is a decimal value between `0` and `1`.

Figure II: Context Aggregation and Question Formulation

Results

- **Performance**:
  - MRI: CNN ~99%, LLMs <60%.
  - Chest CT: CNN ~91%, GPT-4o ~22%, LLaMA ~50%.
  - X-ray: CNN ~83%, GPT-4o ~62%, LLaMA ~65%.
- **Confidence** : CNNs align better; LLMs often **overconfident despite low accuracy**.
- **Efficiency**: CNNs lowest runtime, energy, and CO_2; LLMs significantly higher costs
- **Enhanced filtering**: Boosts GPT-4o accuracy from **62% → 82%**, while reducing runtime and energy





(a) Average Execution Time (b) Average Energy Consumption (c) Average CO_2 Emission

Figure III: Comparative performance of CNNs and LLMs across datasets: Chest X-ray, MRI, and chest CT scan

Limitation

- **LLMs underperform** across modalities, especially CT (e.g., GPT-4o: 22% accuracy).
- **Overconfidence issue**: high confidence despite low accuracy.
- **Method clarity & reproducibility**: filtering pipeline and threshold choice need refinement.
- **High computational cost**: runtime, energy, CO_2 remain barriers for deployment.

Future Work

- **Refine filtering & calibration**: Robust algorithms, reliability metrics, post-hoc scaling.
- **Analyze modality gaps**: why CT Scan is harder than MRI/X-ray.
- **Benchmark vs. multimodal baselines** (CLIP, MedCLIP, CheXzero).
- **Hybrid CNN–LLM frameworks**: combine efficiency + contextual reasoning with explainability.

shibbir@txstate.edu
shahnewazkarim-sakib@utc.edu
adas@uakron.edu

ICIP 2025 Paper

S. Ahmed, S. K. Sakib and A. B. Das, "Can Large Language Models Challenge CNNs in Medical Image Analysis?," 2025 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2025, pp. 2874-2879, doi: 10.1109/ICIP55913.2025.11084721.

TEXAS STATE UNIVERSITY

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM