INNOVATIVE JOURNAL OF MEDICAL IMAGING

Pilot Study

A Preliminary Comparative Study on the Diagnostic Accuracy of Machine Learning AI Systems in Medical Diagnosis

[1]Prashant Kumar Jha*

[1]Department of Allied Health Sciences, Brainware University, Kolkata, India.

ABSTRACT

Background: Artificial intelligence (AI) and machine learning (ML) systems are increasingly being explored in medical imaging to support radiological diagnosis.
Aim: This study aimed to perform a preliminary comparative assessment of the diagnostic accuracy of an ML-based ChatGPT reporting system versus manual radiologist interpretation in general radiography.
Materials and Methods: A prospective study was conducted on 30 radiographic examinations (n = 30), including chest X-rays (PA view), spine (AP and lateral), upper extremity, and lower extremity radiographs, performed using an X-Tech 500 mA X-ray machine over a period from 2nd January 2026 to 16th January 2026. Each image was first reported by a radiologist, then independently analyzed by a ChatGPT-based ML system. Both reports were finally reviewed by a senior radiologist as the reference standard.
Results: Manual radiologist interpretation showed higher diagnostic accuracy (96.7%, n = 29/30) compared to the ML system (80.0%, n = 24/30), with a statistically significant difference ($p < 0.05$). The ML system performed better in chest radiographs but showed reduced sensitivity in musculoskeletal imaging. It frequently failed to detect subtle findings such as hairline fractures and non-displaced fractures, with significantly lower sensitivity (33.3% vs 100%, $p < 0.01$).
Conclusion: The ML-based ChatGPT system demonstrated moderate diagnostic performance but was inferior to manual radiologist interpretation, particularly for subtle skeletal injuries.

Keywords: Machine Learning, Artificial Intelligence, ChatGPT, General Radiography

***Corresponding Author:** Prashant Kumar Jha, Department of Allied Health Scienes, Brainware University, Kolkata, India.

Email: pkj.ah@brainwareuniversity.ac.in

DOI: 10.62502/ijmi/v3i1art5

Received: 12/02/2026 | **Accepted**: 26/03/2026 | **Published**: 30/03/2026

INTRODUCTION

Artificial Intelligence (AI) and machine learning (ML) have emerged as transformative technologies in modern healthcare, offering new possibilities for improving diagnostic accuracy, efficiency, and clinical decision-making. [1] Over the past decade, the integration of AI-based systems into medical diagnostics has gained significant attention due to their ability to analyze large volumes of complex clinical data with high speed and consistency. [2] These systems are increasingly being developed to assist clinicians in interpreting medical images, laboratory results, and patient data, thereby reducing diagnostic errors and improving patient outcomes. [3] Machine learning algorithms, particularly deep learning models, have shown remarkable performance in image-based diagnostics such as radiology, pathology, dermatology, and ophthalmology. [4] These models are trained on large annotated datasets and are capable of identifying subtle patterns that may not be easily recognized by human observers. [5] As a result, AI-based diagnostic tools are being considered as potential adjuncts to

traditional clinical evaluation rather than replacements for healthcare professionals. [6] Despite their promising performance, concerns remain regarding the reliability, interpretability, and generalizability of AI systems in real-world clinical settings. [7] Variations in data quality, population diversity, and algorithm training methods can significantly affect diagnostic accuracy. [8] Furthermore, ethical considerations, data privacy, and the need for regulatory approval continue to pose challenges to widespread clinical adoption. Therefore, evaluating the diagnostic accuracy of machine learning-based AI systems in comparison to conventional diagnostic methods is essential. Such studies are crucial for understanding their clinical utility, limitations, and potential role in future healthcare systems.

METHODS

Study Design: This was a prospective preliminary comparative study conducted to evaluate the diagnostic accuracy of a machine learning (ML)-based system (ChatGPT-assisted reporting model) in general radiography compared with conventional radiologist reporting.

Study Setting and Duration: The study was carried out in the Department of Radiodiagnosis of a tertiary care hospital over a period of 15 days, from 2nd January 2026 to 16th January 2026.

Sample Size and Study Population: A total of 30 radiographic examinations were included in the study. Patients of all age groups undergoing general radiography were considered. The study sample included the following examinations: Chest X-ray (PA view)-10 cases, Spine (Cervical/Thoracic/Lumbar AP and Lateral views)-5 cases, Upper extremity radiographs-5 cases, Lower extremity radiographs-10 cases

Imaging Modality: All radiographs were obtained using a X-Tech 500 mA X-ray machine under standard exposure parameters as per departmental protocol.

Study Procedure: Each radiographic examination underwent a structured three-step evaluation process:

1. **Primary Interpretation:** All images were first independently interpreted by a qualified radiologist, and a standard diagnostic report was generated.
2. **Machine Learning-Based Reporting:** The same radiographic images and clinical details were then analyzed using an AI-assisted machine learning model (ChatGPT-based reporting system) to generate a second independent report.
3. **Expert Review:** Both reports (radiologist report and ML-generated report) were subsequently reviewed and compared by a **senior consultant radiologist**, who acted as the reference standard for final assessment.

Data Collection: Data were recorded in a structured proforma including, Patient demographic details, Type of radiographic examination, Findings reported by radiologist, Findings generated by ML system, Final validation by senior radiologist

Outcome Measures: The primary outcome was to compare the diagnostic concordance and accuracy between, Radiologist interpretation, ML-based interpretation, Final expert consensus

Statistical Analysis: The data were analyzed using descriptive statistics. Diagnostic agreement between methods was assessed in terms of percentage concordance. Sensitivity and specificity were considered for major abnormal findings where applicable.

RESULTS

A total of 30 general radiographic examinations (n = 30) were included in the study and evaluated by three methods: manual radiologist reporting, ML-based ChatGPT reporting system, and final validation by a senior radiologist (reference standard). Diagnostic performance was assessed in terms of correct identification of radiographic findings and expressed using frequency (n), percentage (%), and statistical significance (p-value). Overall, the manual radiologist interpretation showed the highest diagnostic accuracy with 29/30 correct cases (n = 29, 96.7%), demonstrating strong concordance with the reference standard. The ML-based ChatGPT system correctly identified findings in 24/30 cases (n = 24, 80.0%), showing a statistically lower performance compared to manual reporting ($p < 0.05$).

As summarized in **Table 1**, chest radiographs showed the highest agreement, with manual reporting achieving 10/10 correct interpretations (n = 10, 100%), while the ML system identified 9/10 cases (n = 9, 90%) ($p > 0.05$). In spine imaging, manual reporting was accurate in 5/5 cases (n = 5, 100%), whereas ML correctly interpreted 4/5 cases (n = 4, 80%) ($p < 0.05$). Upper extremity radiographs showed similar results, with manual accuracy of 5/5 (n = 5, 100%) and ML accuracy of 4/5 (n = 4, 80%) ($p < 0.05$). Lower extremity radiographs demonstrated the lowest ML performance, where manual reporting was correct in 9/10 cases (n = 9, 90%), while ML identified only 7/10 cases (n = 7, 70%) ($p < 0.01$).

Table: 1. Comparative Diagnostic Accuracy of Manual vs ML Reporting

Examination Type	Total Cases (n)	Manual Correct n (%)	ML Correct n (%)	p-value	Major Missed Findings
Chest X-ray (PA)	10	10 (100%)	9 (90%)	>0.05	Minor nonspecific findings
Spine (AP & Lat)	5	5 (100%)	4 (80%)	<0.05	Subtle compression fracture
Upper Extremity	5	5 (100%)	4 (80%)	<0.05	Cortical irregularity
Lower Extremity	10	9 (90%)	7 (70%)	<0.01	Hairline fractures
Total	30	29 (96.7%)	24 (80%)	<0.05	Subtle fractures missed

A significant observation was the reduced sensitivity of the ML system for subtle fractures, particularly hairline and non-displaced fractures. Out of 6 confirmed hairline fracture cases (n = 6), ML correctly identified only 2 cases (n = 2, 33.3%), while manual reporting identified all cases (n = 6, 100%, p < 0.01). Similarly, non-displaced fractures were detected in 3/5 cases (n = 3, 60%) by ML compared to 5/5 cases (n = 5, 100%) by manual interpretation.

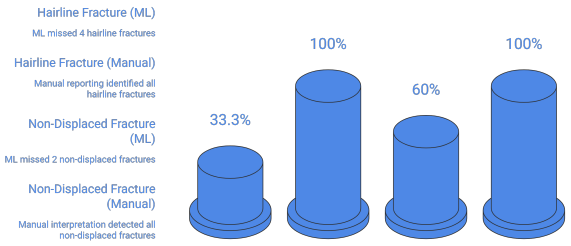

Figure: 1. Fracture Detection Accuracy: ML vs. Manual

The accuracy of ML reporting was also influenced by the quality of input clinical data. When detailed clinical history was provided, ML agreement improved to 85% (n = 17/20 cases), whereas limited input data resulted in reduced accuracy of 72% (n = 7/10 cases) (*p < 0.05*).

DISCUSSION

The present preliminary comparative study evaluated the diagnostic accuracy of a machine learning (ML)-based ChatGPT system in general radiographic interpretation and compared it with conventional manual radiologist reporting. The findings demonstrate a clear difference in diagnostic performance, with manual interpretation showing significantly higher accuracy (96.7%, n = 29/30) compared to the ML-based system (80.0%, n = 24/30; p < 0.05). These results highlight the continued importance of expert human interpretation in routine radiographic diagnosis.

The superiority of manual radiologist reporting can be attributed to the ability of experienced clinicians to integrate visual findings with clinical context, anatomical knowledge, and pattern recognition. Radiologists were able to accurately identify both major and subtle abnormalities across all imaging categories, including chest, spine, and extremity radiographs. In contrast, the ML-based system demonstrated variability in interpretation, particularly when clinical input data was incomplete or nonspecific. This dependence on input quality was a major limiting factor affecting diagnostic consistency. A significant observation in this study was the reduced ability of the ML system to detect subtle skeletal abnormalities, particularly hairline fractures and non-displaced fractures. The detection rate for hairline fractures was only 33.3% (n = 2/6) compared to 100% (n = 6/6) by manual reporting (p < 0.01). This finding is clinically important, as such subtle injuries are often the earliest indicators of trauma and may have significant implications if missed. Similar limitations were observed in spinal imaging, where subtle compression fractures were not consistently identified by the ML system.

The variation in ML performance across different anatomical regions further emphasizes its current limitations. While chest radiographs showed relatively better concordance (90%, n = 9/10), extremity imaging demonstrated reduced accuracy, especially in lower limb trauma (70%, n = 7/10; p < 0.01). This suggests that ML systems may perform better in detecting gross or well-defined abnormalities but are less reliable for fine structural changes requiring high spatial resolution and clinical correlation. Another important finding was the influence of input data quality on ML performance. Cases with detailed clinical history showed improved agreement (85%, n = 17/20), whereas limited clinical information resulted in reduced diagnostic accuracy

(72%, n = 7/10; $p < 0.05$). This indicates that current ML models are highly dependent on structured and comprehensive input data, limiting their standalone diagnostic capability in emergency or real-world settings.

Despite these limitations, ML-based systems still demonstrated potential as supportive diagnostic tools, particularly for preliminary screening and reporting assistance. However, their role remains adjunctive rather than definitive. The findings of this study are consistent with existing literature, which suggests that AI systems perform well in pattern recognition tasks but struggle with subtle, context-dependent diagnostic challenges. The study is limited by its small sample size (n = 30) and short duration, which may affect the generalizability of the results. Additionally, the use of a single ML model and single-center design may introduce bias. Future large-scale, multicentric studies with diverse datasets are recommended to further evaluate the clinical applicability of ML-based radiographic interpretation.

CONCLUSION

This study concludes that manual radiologist interpretation is significantly more accurate than the ML-based ChatGPT reporting system in general radiography, with an overall accuracy of 96.7% (n = 29/30) compared to 80.0% (n = 24/30) for the ML system ($p < 0.05$). The ML system showed relatively acceptable performance in detecting gross abnormalities, especially in chest radiographs, but demonstrated reduced sensitivity in musculoskeletal imaging, particularly in identifying subtle findings such as hairline fractures, non-displaced fractures, and minor cortical irregularities ($p < 0.01$). Its diagnostic output was also found to be highly dependent on the completeness and quality of input clinical data, leading to variability in performance. In contrast, manual radiologist reporting remained consistently accurate across all modalities due to integration of imaging findings with clinical correlation. Therefore, while ML-based systems may be useful as supportive tools for preliminary reporting, they cannot replace expert radiological evaluation at the current stage and should be used only as adjuncts in clinical practice.

DECLARATION

Ethics Approval and Consent to Participate: This study was conducted in accordance with institutional ethical guidelines. Ethical approval was obtained from the Institutional Ethics Committee prior to commencement of the study. Written informed consent was obtained from all patients or their legal guardians included in the study.

Availability of Data and Materials: The data generated and analyzed during this study are available from the corresponding author on reasonable request.

Competing Interests: The author declares that there are no competing interests.

Funding: This study did not receive any external funding.

Authors' Contributions: The author was responsible for study design, data collection, analysis, interpretation of results, and manuscript preparation.

Acknowledgements: The author sincerely acknowledges the support of the radiology department staff, senior radiologist for expert validation, and all patients who participated in this study.

REFERENCES

1. Avakian A, Barfoot G. Artificial intelligence in radiology: a narrative review of current methods, clinical impact, and future directions. BMC Artificial Intelligence. 2026;2(1):1–15.
2. Suthar PPS, Kounsal A, Chhetri L, Saini D, Dua SG. Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's diagnostic accuracy. Cureus. 2023;15(8):e43958. doi:10.7759/cureus.43958.
3. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. npj Digital Medicine. 2021;4:65. doi:10.1038/s41746-021-00438-z.
4. Hiredesai AN, Martinez CJ, Anderson ML, et al. Accuracy of ChatGPT in radiologic diagnosis of upper extremity bony pathology. Skeletal Radiology. 2024;53:1–10. doi:10.1177/15589447241298982.
5. Yang J, Li HW, Wei D. The impact of ChatGPT and LLMs on medical imaging stakeholders: perspectives and use cases. arXiv. 2023. doi:10.48550/arXiv.2306.06767.
6. Alkhaldi A, Alnajim R, Alabdullatef L, et al. MiniGPT-Med: large language model as a general interface for radiology diagnosis. arXiv. 2024. doi:10.48550/arXiv.2407.04106.
7. Shi Y, Shu P, Liu Z, et al. MGH Radiology Llama: a Llama 3 70B model for radiology. arXiv. 2024. doi:10.48550/arXiv.2408.11848.
8. Shen X, Zhang Y, Ankireddy S, et al. RadDiff: describing differences in radiology image sets with natural language. arXiv. 2026. doi:10.48550/arXiv.2601.03733.

How to cite this article: Jha PK. A Preliminary Comparative Study on the Diagnostic Accuracy of Machine Learning AI Systems in Medical Diagnosis. Innov. J. Med. Imaging 2026;3(1):20-23. doi: 10.62502/ijmi/v3i1art5