Journal Homepage: -www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/22893
DOI URL: http://dx.doi.org/10.21474/IJAR01/22893

RESEARCH ARTICLE

CAN ARTIFICIAL INTELLIGENCE REPLACE THE ROLE OF THE RADIOLOGIST IN THE REPORTING OF WRIST RADIOGRAPHS

Nara Evelyn Ng[1], Fred Lam[2] and Francis Lam[3]

1. University of Exeter, UK.
2. Cardiff College, UK.
3. Gleneagles Hospital, Orthopaedic Surgeon, Hong Kong.

Manuscript Info

Manuscript History
Received: 19 December 2025
Final Accepted: 20 January 2026
Published: February 2026

Key words:-
Artificial Intelligence, Orthopaedics, Radiology, Wrist fractures, Wrist abnormalities

Abstract

Purpose: Artificial intelligence (AI) has demonstrated improved accura cy and efficiency in several areas of medical imaging; however, its role in musculoskeletal radiograph interpretation remains unclear.This study aimed to evaluate the diagnostic accuracy of readily accessible AI platf orms in interpreting wrist radiographs and to determine their ability to detect abnormalities and provide correct diagnoses compared with radiologist reports.

Methods:This is a retrospective observational study of 100 consecutive patients who underwent radiographic examination of the wrist referred by the senior author performed at Gleneagles Hospital.A total of 100 anonymised wrist radiographs were included in this studycomprising of 50 normal radiographs and 50 abnormal radiographs, with the abnormal images further categorised as trauma-related, degenerative, congenital, or post-operative. Each radiograph was uploaded to two AI platforms (Grok and CT-read) using a standardised prompt requesting a holistic report. AI outputs were assessed for abnormality detection and diagnost ic correctness. Sensitivity, specificity, accuracy, positive predictive val ue (PPV), negative predictive value (NPV), and Youden's index were calculated and compared against chance performance.

Results: Both AI platforms demonstrated limited diagnostic performance. Grok showed sensitivity above chance for fracture detection but poor specificity, frequently over-identifying abnormalities, with an overall accuracy of 41% and a negative Youden's index. For abnormality detection, Grok achieved moderate accuracy (56%) but limited discriminatory ability. CT-read demonstrated higher specificity than sensitivity, performing better at identifying normal radiographs than detecting abnormalities. Its overall diagnostic accuracy was 55%, with PPV of 0.53 and NPV of 0.64. Across both platforms, performance metrics were near chance levels.

Conclusion: Freely accessible AI platforms showed limited reliability in interpreting wrist radiographs and are not yet suitable for independent clinical decision-making. Further model refinement and training are required before such tools can be safely integrated into musculoskeletal imaging practice.

Corresponding Author:- Nara Evelyn Ng
Address:- University of Exeter,UK.

Introduction:-

The use of artificial intelligence (AI) is increasingly common in medicine. The quoted advantages of AI generated radiograph reporting include improved detection and highlighting of potential abnormalities on radiographs, guiding the clinician's attention to areas that might otherwise be overlooked, improved sensitivity, increased efficiency with reduction in the time it takes for a radiologist to interpret an image, increased overall productivity, reducing diagnostic errors and lastly a potential source for support and teaching for trainees. Furthermore, in many clinics and hospitals worldwide, an instantly available radiology report is not always possible especially when the radiographic examination is performed outside normal working hours. Typically, the radiologist report is available within a few days of the examination, and the report is then sent back to the referring clinician. The report is then checked and if there has been a missed diagnosis, the patient is contacted and brought back for further treatment. This can adversely affect the patient's experience.

Bone fractures stand out as a critical area where instantly available reports can greatly help the referring clinician in making the correct diagnosis and instigating early accurate treatment. The use of AI in chest radiography reporting is already widespread, and studies have shown that there is an absolute increase in accuracy and reduction in reading times for radiologists of all levels of experience. However, little is known regarding the accuracy of AI in interpreting musculoskeletal radiographs. We chose to study wrist radiographs only in this study as this is one of the most common radiographs performed and it covers both traumatic and non-traumatic conditions (congenital abnormalities, degenerative conditions, post-surgical complications).The aim is to evaluate the diagnostic accuracy (sensitivity, specificity, positive predictive value, negative predictive value) of AI in interpreting wrist radiographs. The study would also aim to identify the type of conditions (e.g. bone fractures) which some studies claim AI has high accuracy, as well as some conditions (e.g.trauma-related, degenerative, congenital, or post-operative) in which models and algorithms have not yet been established.

Method:-

Ethical Approval:-

This retrospective study was reviewed and approved by the University of Hong Kong - Gleneagles Hospital Hong Kong Institutional Review Board. The requirement for informed consent was waived due to the retrospective nature of the study and the use of anonymized radiographic and clinical data. The study was conducted in accordance with the Declaration of Helsinki and relevant local regulations.

Materials and Design:-

This is a retrospective observational study of 100 consecutive patients who underwent radiographic examination of the wrist referred by the senior author performed at Gleneagles Hospital. A total of 100 anonymised wrist radiographs were included in this study comprising of 50 normal radiographs and 50 abnormal radiographs, with the abnormal images further categorised as trauma-related, degenerative, congenital, or post-operative. The inclusion criteria include all patients who underwent radiographic posterior-anterior and lateral view examination referred from outpatient and inpatient settings as well as accident and emergency. Wrist radiographs were retrieved from the PACS database of Gleneagles Hospital Hong Kong. All images were obtained as part of routine clinical practice following standardized institutional imaging protocols. Radiographs were screened for quality prior to analysis and were included only if they demonstrated adequate exposure, correct positioning, and complete visualization of the relevant anatomical landmarks. Images with motion artefact, poor positioning, or incomplete views were excluded.

Procedure:-

All eligible radiographs were collected and anonymised prior to analysis. Each image was reviewed and confirmed to meet inclusion criteria before being uploaded to the AI platforms (Grok and CT-read). The uploads were performed sequentially to ensure identical handling across platforms.Each radiograph was individually uploaded to the selected AI platforms. To maintain consistency, the following standardised prompt was used for all platforms: "Can you provide a holistic report for the uploaded x-ray?". In this study, the definition of a correct diagnosis is based on whether it accurately states the location and the type of abnormality that is present in the x-ray, in which it would be compared to the opinion of a radiologist – For example, if a fracture is detected but the anatomical site is identified incorrectly, the report is classified as incorrect.

The AI-generated responses were recorded verbatim. Each AI report was analysed to determine:
1. Whether the AI correctly identified the presence or absence of a fracture, and
2. Whether the AI recognized any other radiographic abnormality or notable finding.
The diagnostic output for each image was then categorised using two binary variables:
• Abnormality present: Yes or No
• Diagnosis correct: Yes or No

To ensure diagnostic validity and establish a reliable reference standard, the senior author, blinded to all AI-generated outputs, independently reviewed every radiograph alongside its corresponding hospital radiologist report. This process served to confirm concordance between the imaging findings and the documented diagnoses. Where discrepancies arose between the radiologist report and the senior author's interpretation, consensus was reached through collaborative review of the imaging findings and relevant clinical documentation. Once verified, the radiologist reports were adopted as the gold standard against which all AI diagnostic outputs were compared.

Statistical Analysis:-
Diagnostic performance metrics including sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), likelihood ratios, and Youden's index were calculated for each AI platform.To determine whether performance exceeded random classification, sensitivity and specificity were tested against a null hypothesis value of 0.5, representing chance-level discrimination in a binary diagnostic task (normal vsabnormal). A value significantly greater than 0.5 indicates performance superior to random guessing, whereas values significantly below 0.5 indicate systematic misclassification. Statistical significance was set at $p < 0.05$.

Results:-

All analyses were conducted using Jamovi v2.4.12.0
The main analysis assessed the diagnostic performance of Grok and CT-read on wrist radiographs. Metrics evaluated included sensitivity, specificity, PPV, NPV, accuracy, and Youden's index (J). Observed values were compared against chance ($H_0 = 0.50$) to determine each platform's ability to detect abnormalities. Results are presented in Tables 1–8.

Table 1:
Sensitivity and specificity for diagnostic ability of Grok:-

Support: Diagnostic statistics

	Value	Difference	S	Param	G	df	p
H_0 vs Sensitivity	0.5000	0.316	-10.0952	1, 2	21.190	1	< .001
H_0 vs Specificity	0.5000	-0.128	-1.175	1, 2	3.351	1	0.067

Note. S uses Occam's Bonus correction for parameters (Param).

The results showed that sensitivity was significantly higher than chance ($p < .001$), indicating that the AI platforms could reliably detect fractures when present. However, specificity did not differ significantly from chance ($p = 0.067$), suggesting limited accuracy in identifying normal radiographs

Table 2:
Accuracy and NPV and PPV for diagnostic ability of Grok:-

Diagnostic statistics

Statistic	LR	Neg LR	Accuracy	Odds Ratio	Prevalence	PPV	NPV	Youden's index J
Value	0.493	1.301	0.410	0.379	0.490	0.321	0.444	-0.1889

The results showed an overall diagnostic accuracy of 41%, with a positive predictive value (PPV) of 32% and a negative predictive value (NPV) of 44%. The negative Youden's index (J = –0.19) indicates discriminative ability well below acceptable clinical thresholds. The low PPV suggests a high rate of false positives, meaning the platform lacks clinical utility for confirming fracture presence.

Table 3:
Sensitivity and specificity results for abnormality detection of Grok

Support: Diagnostic statistics						
	Value	**Difference**	**S**	**Param**	**G**	**df**	**p**
H_0 vs Sensitivity	0.5000	0.0556	0.2216	1, 2	0.5567	1	0.456
H_0 vs Specificity	0.5000	-0.1545	-2.1708	1, 2	5.3416	1	0.021
Note. S uses Occam's Bonus correction for parameters (Param).							

The results showed that specificity was significantly below chance (p = 0.021), while sensitivity did not differ significantly from chance (p = 0.456). These findings suggest that the model tended to over-identify abnormalities and lacked consistent discrimination between normal and abnormal images.

Table 4:
Accuracy and NPV and PPV for abnormality detection of Grok

Diagnostic statistics								
Statistic	**LR**	**Neg LR**	**Accuracy**	**Odds Ratio**	**Prevalence**	**PPV**	**NPV**	**Youden's index J**
Value	1.287	0.849	0.560	1.516	0.450	0.513	0.590	0.099

The results showed a moderate overall accuracy of 56%, with PPV = 0.51 and NPV = 0.59. The slightly positive Youden's index (J = 0.10) suggested marginal improvement over chance performance but still reflected limited diagnostic accuracy for detecting abnormalities.

Table 5:
Sensitivity and specificity for diagnostic ability of CT-read

Support: Diagnostic statistics							
	Value	**Difference**	**S**	**Param**	**G**	**df**	**p**
H_0 vs Sensitivity	0.500	-0.337	-11.657	1, 2	24.314	1	< .001
H_0 vs Specificity	0.500	0.226	-4.878	1, 2	10.756	1	0.001
Note. S uses Occam's Bonus correction for parameters (Param).							

The results showed significant differences from chance for both sensitivity (p < .001) and specificity (p = 0.001). Sensitivity was lower than 0.5, whereas specificity exceeded chance levels, indicating that the platform was more effective at identifying normal images than detecting abnormal findings.

Table 6:
Accuracy and NPV and PPV for diagnostic ability of CT-read

Diagnostic statistics								
Statistic	**LR**	**Neg LR**	**Accuracy**	**Odds Ratio**	**Prevalence**	**PPV**	**NPV**	**Youden's index J**
Value	1.153	0.595	0.550	1.939	0.490	0.526	0.636	0.111

The results showed an overall diagnostic accuracy of 55%, with a PPV of 53% and an NPV of 64%. The Youden's index (J = 0.11) reflects marginal discriminative ability. While the NPV suggests a modest capacity to rule out normal cases, the overall diagnostic capability is insufficient to support standalone clinical decision-making.

Table 7:
Sensitivity and specificity results for abnormality detection of CT-read

Support: Diagnostic statistics						
	Value	**Difference**	**S**	**Param**	**G**	**df**	**p**
H_0 vs Sensitivity	0.500	-0.189	-2.792	1, 2	6.584	1	0.010
H_0 vs Specificity	0.500	0.246	-6.423	1, 2	13.846	1	< .001
Note. S uses Occam's Bonus correction for parameters (Param).							

The results showed statistically significant sensitivity (p = 0.010) and specificity (p < .001). The higher specificity values indicated that the system performed better at correctly identifying normal wrist radiographs than detecting abnormal cases.

Table 8:
Accuracy and NPV and PPV for abnormality detection of CT-read

Diagnostic statistics								
Statistic	**LR**	**Neg LR**	**Accuracy**	**Odds Ratio**	**Prevalence**	**PPV**	**NPV**	**Youden's index J**
Value	0.924	1.222	0.450	0.7561	0.450	0.431	0.500	-0.057

The results showed an overall accuracy of 45%, with a PPV of 43% and an NPV of 50%. The Youden's index (J = – 0.06) is effectively zero, indicating that the platform performs no better than random guessing. This level of performance renders the test clinically useless, as it fails to provide any meaningful discrimination between normal and abnormal findings.

Summary:-
Overall, the results showed that both Grok and CT-read demonstrated limited diagnostic performance on wrist radiographs. Sensitivity and specificity varied across tasks, with specificity generally outperforming sensitivity, indicating that the platforms were more reliable at identifying normal images than detecting abnormalities. Accuracy, positive predictive value (PPV), and negative predictive value (NPV) were moderate too low for both systems, and Youden's indices were near zero or slightly negative. These findings suggest that, while the platforms performed slightly better than chance in some measures, their overall reliability for interpreting wrist radiographs was limited. Collectively, these results highlight the challenges of using freely accessible AI tools for clinical imaging interpretation and set the stage for further discussion regarding their potential limitations and areas for improvement. The frequently quoted benefits for AI generated radiology report is that it can significantly reduce the radiologist time and enhanced efficiency as well as improved accuracy. However, our study has shown that the accuracy rate of commonly available AI tools is only approximately 50/50. If as many as half the radiology reports must be edited manually by the radiologist, it creates a dilemma whether if AI should be considered in this field.

Discussion:-

Artificial intelligence has been widely promoted as a tool that can enhance clinical efficiency, highlight subtle abnormalities, and support clinicians when immediate radiology input is not available[1]. This is particularly relevant in musculoskeletal imaging, where delayed or missed fracture detection can negatively impact patient outcomes. Although AI has demonstrated strong performance in other areas such as chest radiography, evidence shows that AI systems can detect up to 90% of pulmonary nodules and can improve radiologist sensitivity by approximately 9 to 10 percent[2]. In contrast, much less is known about its performance in interpreting wrist radiographs across both traumatic and non-traumatic presentations. Recent studies have focused mainly on fracture identification[3], and there remains a lack of research exploring broader musculoskeletal applications. Our study aimed to address this gap by evaluating the diagnostic accuracy of two AI systems, Grok and CT-read.Importantly, the AI systems evaluated in this study were not purpose-built medical imaging models, but general-purpose multimodal platforms applied in a zero-shot manner without radiology-specific training or calibration. This distinction is critical when comparing our findings with prior literature. Many previously published musculoskeletal AI systems were developed using supervised deep learning architectures trained on large, annotated radiographic datasets and optimized for fracture detection. In contrast, general-domain models are increasingly accessible to clinicians and the public despite lacking regulatory approval, domain validation, or standardized performance benchmarks.

From a clinical workflow perspective, the widespread availability of such non-medical AI tools raises important concerns. Clinicians may encounter AI-generated interpretations outside formal radiology systems, particularly in settings with limited specialist access. However, the near chance-level performance observed in this study suggests that reliance on these tools for primary interpretation would introduce substantial diagnostic risk. At best, such systems may function as informal adjuncts requiring expert oversight rather than independent decision-making tools.These findings also highlight the importance of domain-specific training and regulatory oversight in the development of AI for musculoskeletal imaging. Improved performance will likely require targeted fine-tuning on large, annotated radiographic datasets, standardized external validation, and prospective evaluation within real-world clinical workflows. Future research should clearly distinguish between regulated diagnostic systems and general-purpose AI platforms repurposed for clinical interpretation.Although this study contributes to the limited literature on AI interpretation of wrist radiographs using publicly accessible platforms, several important limitations should be acknowledged.

First, the sample size was relatively small (n = 100). This limited sample may not adequately represent the full spectrum of fracture types, severity, and anatomical variations encountered in clinical practice, potentially affecting the generalizability of the findings. A larger cohort would be necessary to obtain more precise estimates of sensitivity, specificity, and predictive values.Second, the retrospective design introduces inherent biases. While a licensed radiologist established the ground truth for all 100 wrist radiographs, the retrospective nature of the data collection means the images were not acquired under controlled or standardized conditions for AI interpretation. This design may not fully capture the variability of real-time clinical decision-making, and the controlled reading environment differs from the pressures of a live clinical workflow. Prospective validation is required to determine how these AI tools would perform in real-time clinical settings.Lastly, an important limitation is that it relies on non–medical-grade AI applications (Grok and CT-scan). These systems differ substantially from the clinically validated AI tools typically available in large hospitals and academic centres. Unlike their medical-grade counterparts, the platforms evaluated here were not designed for diagnostic purposes; they lack rigorous validation on curated medical datasets and have not undergone regulatory review for clinical safety or efficacy. Therefore, while our findings offer insight into the capabilities of publicly accessible AI, they cannot be generalized to specialized clinical AI systems, and the use of such non-medical-grade tools for actual patient care would be inappropriate without further prospective validation.

However, this limitation also underscores the study's relevance to real-world practice. The clinicians most likely to depend on AI for immediate radiograph interpretation, are often the ones working in rural clinics or urgent care centres, or low-resource settings, in which access to expensive, medical-grade platforms is limited. Understanding how readily available, publicly accessible AI models perform under these conditions is therefore essential, even if their diagnostic capabilities are not directly comparable to specialized clinical systems. These findings should be interpreted as exploratory and hypothesis-generating rather than as definitive evidence of clinical efficacy.In the present study, both AI platforms exhibited variable performance, with several metrics approaching chance-level discrimination. Grok demonstrated sensitivity for fracture detection that exceeded random classification, but its

specificity did not differ significantly from chance. In contrast, CT-read showed higher specificity but reduced sensitivity, resulting in an inconsistent diagnostic profile. Across broader abnormality detection tasks, overall accuracy, predictive values, and Youden's index remained modest, with many measures approximating chance-level performance ($H_0 = 0.5$).The results showed that neither system currently performs at a level that would reliably support clinical decision making. Grok demonstrated inconsistent diagnostic behaviour and frequently identified abnormalities where none were present. Although its sensitivity for distal radius fractures exceeded chance, the low specificity and below chance overall performance suggest that it may generate unnecessary follow up and patient anxiety. Its difficulty distinguishing normal from abnormal wrist radiographs, particularly in the abnormality detection condition, indicates that it is not yet suitable for routine musculoskeletal interpretation.CT-read performed relatively better, especially in recognising normal radiographs, which is important for reducing unnecessary referrals. However, its low sensitivity to fractures means that clinically important injuries could still be missed. Since early and accurate fracture identification is one of the main reasons for incorporating AI into frontline practice, this limitation significantly restricts its usefulness.Overall, these findings suggest that although AI has the potential to improve workflow and support clinicians who do not have immediate access to radiology reporting, current musculoskeletal models are not yet ready for clinical deployment. More extensive training, greater dataset diversity, and continued refinement are needed before these tools can reliably assist in the interpretation of wrist radiographs.

Bibliography:-

1. R. Najjar, "Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging," Diagnostics, vol. 13, no. 17, p. 2760, Aug. 2023, doi: 10.3390/diagnostics13172760.
2. M. Meetschenet al., "AI-Assisted X-ray Fracture Detection in Residency Training: Evaluation in Pediatric and Adult Trauma Patients," Diagnostics, vol. 14, no. 6, p. 596, Mar. 2024, doi: 10.3390/diagnostics14060596.
3. M. E. Adam Essa, "Diagnostic accuracy of AI in chest radiography for pneumonia and lung cancer: A meta-analysis," Eur. J. Radiol. Open, vol. 15, p. 100701, Dec. 2025, doi: 10.1016/j.ejro.2025.100701.