Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-Assisted Failure Prediction on Postoperative X-Rays: A Head-to-Head Comparison with Consultant Trauma Surgeons

Gaurav Vatsa

Vatsagaurav07@gmail.com

Viraat Ramayan Institute of Medical Sciences

Dhaval Shah

B Lal Hospital and Research Centre

Saurabh Suman

All India Institute of Medical Sciences, Deoghar

Article

Keywords: Artificial Intelligence, Convolutional Neural Network, Postoperative Radiography, Failure Prediction, Trauma Surgery, Diagnostic Accuracy, Orthopaedic Imaging

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11408305/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background & Objectives:

Surgical hardware failure after orthopaedic fracture fixation causes morbidity and rising cost. Immediate postoperative radiographs may carry subtle markers of mechanical failure. This study compared a convolutional neural network-based artificial intelligence model with three blinded trauma surgeons in predicting 12-month fixation failure and assessed the model's potential as a clinical decision-support adjunct.

Material & Methods

This prospective multicentre study analysed 466 patients (5–91 years; 59% male; 30% open fractures) with operatively treated humerus, femur, ulna, radius or tibia fractures and standardised immediate postoperative radiographs. Failure was a composite endpoint of hardware breakage, screw or plate migration, non-union, or malunion requiring revision within 12 months, adjudicated independently. A ResNet-50 CNN, fine-tuned on an independent 200-patient development cohort with no overlap with the 466-patient application cohort, was compared with three surgeons using sensitivity, specificity, positive predictive value, negative predictive value, accuracy, and class-wise F1 score, with paired testing via McNemar's test.

Results

Of 466 patients, 104 had fixation failure (22.3%). The AI model achieved sensitivity 96.15%, specificity 98.90%, PPV 96.15%, NPV 98.90%, F1 (failure class) 96.15%, and accuracy 98.28% (458 of 466 correct). The strongest surgeon achieved sensitivity 91.35%, specificity 98.62%, PPV 95.00%, F1 (failure class) 93.14%, and accuracy 97.00% (452 correct). The AI significantly outperformed Surgeons 1 and 2 (exact binomial $p < 0.0001$ and $p = 0.0001$). The comparison with Surgeon 3 did not reach statistical significance (exact binomial $p = 0.0703$).

Conclusions

The AI model outperformed every participating trauma surgeon across all metrics and significantly outperformed two of three surgeons; its advantage over the most experienced surgeon did not reach statistical significance in this cohort. These findings support a possible role as a clinical decision-support adjunct rather than a clinician replacement, pending external prospective validation, subgroup analysis by anatomical site and fixation technique, and resolution of ambiguity in the composite failure endpoint used to adjudicate the primary outcome.

Introduction

Surgical hardware failure remains a major clinical problem, leading to patient complications, repeat surgeries and increased healthcare costs[1]. These failures often result from poor implant positioning or construct instability, emphasizing the need for early detection beyond routine clinical review. Immediate postoperative radiographs show the final bone–implant construct and therefore provide a valuable basis for deep-learning models designed to detect subtle fixation problems and predict mechanical failure[2]. We compare an AI-driven prognostic model against human expert assessments to address the gap in validated clinical decision tools for orthopaedic trauma. As diagnostic algorithms become more widespread, performance validation is essential to bridge the divide between published predictive models and routine clinical practice[3]. We evaluate whether machine learning can reliably identify early markers of failure that humans might miss, helping to clarify the role of AI in reducing patient harm[4]. This research acknowledges the ongoing debate regarding the superiority of machine learning over conventional statistical methods in forecasting binary adverse clinical outcomes[5]. Furthermore, while clinical adoption of such technology is viewed as inevitable, concerns regarding the reliability of algorithmic predictions necessitate direct benchmarking against the current gold standard of surgical expertise[6].

Deep learning has shown it works well for detecting and classifying fractures, especially when networks are trained for specific orthopaedic tasks[7]. Researchers have successfully used these systems to diagnose ankle, wrist, and knee fractures[8], often achieving over 90% accuracy in standard tests[9]. That said, existing tools often prioritize just spotting a fracture, rather than tracking the long-term risks seen in complex postoperative cases[10]. Recent studies are moving from simple classification to predicting outcomes, like patient recovery or mechanical complications, by combining images with clinical data[11,12]. Newer neural networks can now spot subtle patterns that predict if a fracture loses alignment which marks a step beyond early studies[13]. Even with these strides, many models still lack the real-world testing needed to prove they work well across different hospitals[3].

Our main goal is to measure how well a ResNet-50–based convolutional neural network predicts early fixation failure using immediate postoperative X-rays. We also compared the model against the assessments of three trauma surgeons to determine whether AI can help sharpen clinical judgment and identify subtle markers of hardware failure that might reduce disagreement between clinicians in complex trauma cases[14,15]. We hypothesize that the AI model will identify high-risk fractures as accurately as experienced trauma surgeons by detecting patterns that point to mechanical instability. By using deep learning to find these patterns, the model aims to reduce the guesswork inherent in human interpretation and improve how we predict patient outcomes. This implementation reflects the growing utility of deep learning in orthopaedics to process complex imaging data and support clinicians in high-stakes decision-making[16,23,26–29].

Methodology

Study design and population

We conducted this prospective study across three trauma centres, enrolling 500 patients in total over the 1-year study period, with ethics approval at each site. We included patients aged 5–91 years undergoing operative fixation of humerus, femur, radius, ulna or tibia fractures (fractures of the appendicular skeleton) with both open and closed injuries permitted, an immediate postoperative X-ray, and a documented 12-month follow-up. We required standardized X-ray protocols at all sites. For patients with fractures at more than one anatomical site, one fracture site was selected at random for imaging and evaluation; surgeons and the AI model evaluated the same randomly-selected image per patient. Three board-certified trauma surgeons independently evaluated these images, remaining blinded to both patients' long-term outcomes and the AI model's predictions. We defined the index examination as the primary postoperative radiograph taken within 24 hours of fracture fixation, using standard anteroposterior and lateral projections. We de-identified all patient data. Images were normalized for resolution and contrast prior to processing, and images with poor bone visibility, rotation, or hardware obscuring the fracture site were excluded to reduce noise[19].

Inclusion criteria

patients aged 5–91 years, with a mix of open and closed limb fractures of the humerus, femur, ulna, radius and tibia; usable postoperative radiographs within 24 hours of fixation; complete 12-month follow-up; informed consent per institutional ethics requirements, including parental consent and child assent for minors[20].

Exclusion criteria

pathological fractures due to malignancy or metabolic bone disease; complex polytrauma (chest, abdominal, or head injury) precluding standardized follow-up[21].

Outcome definition

Postoperative fixation failure was defined as a composite endpoint consisting of symptomatic hardware breakage, screw or plate migration, non-union, or malunion. Non-union was defined as the absence of progressive bridging callus at 6-month follow-up, corroborated by persistent weight-bearing pain[22]. A malunion is broadly defined as a fracture that has healed in a clinically unacceptable position, resulting in deformity or dysfunction[22]. Objective imaging that confirmed physical failure of the implant with clear structural fracture or snapping of plates, rods, wires, or screws, bending or physical loss of implant integrity without necessarily causing gross displacement and migration or secondary displacement of broken fragments into adjacent joint spaces or soft tissues were defined as Implant Failure. All failure outcomes were adjudicated independently by two board-certified orthopaedic trauma surgeons not involved in the study evaluation. The adjudicators were blinded to the evaluators' classifications and to the AI model's output. Patients were followed at 6 weeks, 3 months, 6 months, and 12 months, with any revision surgery during this period documented.

AI model development

Postoperative radiographs were classified using a ResNet-50 convolutional neural network initialized with ImageNet-pretrained weights. Images were resized to 224 × 224 pixels and intensity-normalized before input. The network was fine-tuned on an independent development cohort of 200 patients (one image per patient; for patients with fractures at multiple anatomical sites, one fracture site was selected at random for imaging, without deliberate matching to the site of eventual failure outcome). The development cohort was split at the patient level into training (n = 100), validation (n = 50), and test (n = 50) sets, stratified by failure status. This development cohort had no patient- or image-level overlap with the 466-patient evaluable application cohort used for the AI-versus-surgeon comparison reported in Results (500 patients were enrolled, of whom 34 were lost to follow-up and excluded). The development cohort was drawn from the same three centres as the application cohort. The network was fine-tuned using the AdamW optimizer with an initial learning rate of 1×10^{-4}, with binary cross-entropy with logits as the loss function. Training was performed for a maximum of 100 epochs with early stopping based on validation performance; the checkpoint achieving the best validation AUROC was retained for final evaluation. Data augmentation (small random rotations (± 5°), translations (± 5%), scaling (0.95–1.05×), and mild brightness/contrast variation) was applied exclusively to the training set; no augmentation was applied to the validation or test sets. Model output was binarized using a fixed, pre-specified default threshold of 0.5, selected prior to any evaluation to avoid post hoc threshold tuning. All preprocessing, augmentation, and thresholding parameters were fixed prior to evaluation on the 466-patient application cohort. To facilitate comparative analysis, all four evaluators (three surgeons and the AI model) independently classified the same 466-patient application cohort.

Statistical analysis

Diagnostic performance was assessed using sensitivity, specificity, PPV, NPV, accuracy, and class-wise F1 score (failure class and success class, plus macro-F1), with fixation failure treated as the positive class. Paired comparisons between the AI model and each surgeon were performed using McNemar's test on patient-level discordant pairs, with continuity-corrected χ^2 and exact binomial p-values reported[17]. Inter-observer agreement among the three surgeons was assessed using κ statistics[18]. Statistical significance was set at $\alpha = 0.05$; given three pairwise AI-versus-surgeon comparisons, a Holm-Bonferroni correction for multiple comparisons was also applied and is reported alongside the uncorrected p-values.

Results

A total of 500 patients were identified or enrolled during the 1-year study period. Thirty-four patients were lost to follow-up and excluded, leaving a final evaluable cohort of 466 patients. Patient ages ranged from 5 to 91 years. Males comprised 59%(275) of the cohort and females 41%(191). Open fractures accounted for 30%(140) of all cases. Fracture fixation was performed using intramedullary nails, plates, screws, titanium elastic nails, cannulated cancellous screws, and Kirschner wires. Of the 466 patients in

the final analysis, 104 experienced postoperative fixation failure and 362 did not (failure prevalence 22.3%). Seventy-three of these 104 patients underwent revision surgery; the remainder met radiographic or clinical criteria for failure but declined further surgery. (Table 1&2)

Table 1
Baseline characteristics of the evaluable cohort

Characteristic	**Value**
Total evaluable patients	466
Age range, years	5–91
Sex — male	275(59%)
Sex — female	191(41%)
Open fractures	140(30%)
Closed fractures	326(70%)
Fracture site — humerus / femur / tibia / radius / ulna	82(17.59%) / 99(21.24%) / 121(25.96%) / 79(16.95%) / 85(18.24%)
Lost to follow-up	34

Table 2
Reported failure outcomes

Failure outcome	**Number of patients**
Implant breakage	21
Screw or plate migration	9
Nonunion	46
Malunion	28
Total	**104**

95% confidence intervals computed from the counts in Table 3 using the Wilson score interval for sensitivity, specificity, PPV, NPV, and accuracy, and a percentile bootstrap for F1 (failure class).

Table 3
Diagnostic performance metrics among the surgeons and AI

Measure	Surgeon 1	Surgeon 2	Surgeon 3	AI Model
True positive	91	93	95	100
False positive	22	15	5	4
True negative	340	347	357	358
False negative	13	11	9	4
Sensitivity	87.50%	89.42%	91.35%	96.15%
Specificity	93.92%	95.86%	98.62%	98.90%
Positive predictive value	80.53%	86.11%	95.00%	96.15%
Negative predictive value	96.32%	96.93%	97.54%	98.90%
F1 score (failure class)	83.87%	87.74%	93.14%	96.15%
F1 score (success class)	95.10%	96.39%	98.08%	98.90%
Macro-F1	89.49%	92.06%	95.61%	97.52%
Accuracy	92.49%	94.42%	97.00%	98.28%

Surgeon 1 recorded 91 true-positive, 22 false-positive, 340 true-negative, and 13 false-negative classifications (sensitivity 87.50%, specificity 93.92%, PPV 80.53%, NPV 96.32%, accuracy 92.49%). Surgeon 2 recorded 93 true-positive, 15 false-positive, 347 true-negative, and 11 false-negative classifications (sensitivity 89.42%, specificity 95.86%, PPV 86.11%, NPV 96.93%, accuracy 94.42%). Surgeon 3 recorded 95 true-positive, 5 false-positive, 357 true-negative, and 9 false-negative classifications (sensitivity 91.35%, specificity 98.62%, PPV 95.00%, NPV 97.54%, accuracy 97.00%). The AI model recorded 100 true-positive, 4 false-positive, 358 true-negative, and 4 false-negative classifications (sensitivity 96.15%, specificity 98.90%, PPV 96.15%, NPV 98.90%, accuracy 98.28%). The AI model achieved the highest sensitivity, specificity, PPV, NPV, F1 (both classes), and accuracy among all four evaluators. Surgeon 3 was the strongest-performing surgeon, followed by Surgeon 2 and Surgeon 1. The AI model correctly classified 458 of 466 cases, compared with 452 for Surgeon 3, 440 for Surgeon 2, and 431 for Surgeon 1. (Table 4)

Table 4
95% Confidence Intervals for Diagnostic Performance Metrics

Metric	Surgeon 1	Surgeon 2	Surgeon 3	AI Model
Sensitivity, % (95% CI)	87.50 (79.78–92.55)	89.42 (82.05–93.99)	91.35 (84.37–95.38)	96.15 (90.53–98.49)
Specificity, % (95% CI)	93.92 (90.97–95.95)	95.86 (93.28–97.47)	98.62 (96.81–99.41)	98.90 (97.19–99.57)
PPV, % (95% CI)	80.53 (72.28–86.78)	86.11 (78.34–91.40)	95.00 (88.82–97.85)	96.15 (90.53–98.49)
NPV, % (95% CI)	96.32 (93.80–97.84)	96.93 (94.58–98.28)	97.54 (95.39–98.70)	98.90 (97.19–99.57)
F1, failure class, % (95% CI)	83.87 (78.26–88.79)	87.74 (82.80–92.12)	93.14 (89.20–96.43)	96.15 (93.20–98.55)
Accuracy, % (95% CI)	92.49 (89.73–94.55)	94.42 (91.95–96.16)	97.00 (95.02–98.20)	98.28 (96.65–99.13)

Paired Comparison of AI and Surgeon Classifications

McNemar's test, reconstructed directly from patient-level predictions, showed the AI model significantly outperformed Surgeon 1 (29 discordant pairs favoring the AI vs. 2 favoring the surgeon; $\chi^2 = 21.806$, exact binomial $p < 0.0001$) and Surgeon 2 (20 vs. 2; $\chi^2 = 13.136$, exact binomial $p = 0.0001$). The comparison with Surgeon 3 (7 discordant pairs favoring the AI vs. 1 favoring the surgeon; $\chi^2 = 3.125$, exact binomial $p = 0.0703$) did not reach statistical significance at $\alpha = 0.05$.

Because three pairwise AI-versus-surgeon comparisons were performed, exact two-sided binomial p-values were recomputed from the discordant-pair counts in Table 5 and adjusted using the Holm-Bonferroni method. The AI model remained significantly superior to Surgeon 1 (exact $p = 4.6\times10^{-7}$; Holm-adjusted $p = 1.4\times10^{-6}$) and Surgeon 2 (exact $p = 0.000121$; Holm-adjusted $p = 0.000242$). The comparison with Surgeon 3 remained non-significant (exact $p = 0.0703$; Holm-adjusted $p = 0.0703$).

Table 5
Pairwise Discordance Analysis: AI Model vs. Each Trauma Surgeon

Comparison	Discordant favouring AI	Discordant favouring surgeon	χ^2 (corrected)	Exact binomial p
Surgeon 1 vs. AI	29	2	21.806	< 0.0001
Surgeon 2 vs. AI	20	2	13.136	0.0001
Surgeon 3 vs. AI	7	1	3.125	0.0703

Discussion

The principal findings of this study demonstrate that the AI model outperformed all three participating trauma surgeons across every diagnostic performance metric, achieving a sensitivity of 96.15%, a specificity of 98.90%, and an overall accuracy of 98.28%. Its advantage was large and statistically significant against Surgeons 1 and 2, but comparatively narrow and not statistically significant against Surgeon 3. We report class-wise F1 alongside aggregate accuracy for this reason: F1 for the failure class ranged from 83.87% to 96.15%, a wider spread than accuracy alone conveys, since aggregate accuracy is inflated by the larger non-failure class. Because PPV and NPV are prevalence-dependent, the point estimates in Table 3 (AI PPV 96.15%, NPV 98.90%) apply to cohorts with a failure prevalence comparable to the 22.3% observed here and will not transfer directly to settings with different failure rates.

It is also important to situate these results within the broader, multifactorial etiology of postoperative fixation failure. Mechanical factors like implant positioning, construct stiffness and interfragmentary strain are directly visible on postoperative radiographs and fall within the AI model's feature space. Biological determinants of healing, including bone quality, vascularity, soft-tissue injury, infection, and patient factors such as diabetes and smoking status, are not reliably captured by a single immediate postoperative image[24]. In this cohort, non-union (46/104, 44.2%) and malunion (28/104, 26.9%) together account for 74 of 104 failures (71.2%), with implant breakage (21/104) and screw/plate migration (9/104) accounting for the remainder. Because non-union and malunion arise from a combination of mechanical and biological factors, and because they represent the clear majority of the composite endpoint in this cohort, a model trained exclusively on radiographic features may be systematically less sensitive to the specific failure modes that make up most of the outcome being predicted. A decomposed analysis separating mechanically-driven failure (implant breakage, migration) from biologically-influenced failure (non-union, malunion) would substantially strengthen this manuscript; patient-level failure-subtype labels needed for such a decomposed performance analysis were not collected in this cohort, and we acknowledge this as a limitation rather than a planned addition.

The AI model was fine-tuned on an independent 200-patient development cohort (100 training / 50 validation / 50 test, split and stratified at the patient level) with no patient- or image-level overlap with the 466-patient application cohort used for the head-to-head comparison reported here, addressing a central concern about potential optimism bias from data leakage. We note, however, that a training set of 100 images is small relative to the parameter count of a ResNet-50 architecture, even with ImageNet pretraining and augmentation; while performance on the fully independent 466-patient application cohort is reassuring evidence against overfitting to the small development set, this should be stated explicitly as a limitation, and future work should evaluate whether performance is stable when the model is retrained on a larger, independently sourced development cohort. A related and more consequential limitation concerns patients with fractures at more than one anatomical site: random site selection (one fracture site imaged per patient, chosen at random when more than one was present) was used in both the 200-patient development cohort and the 466-patient application cohort. Provided the same randomly-selected image was shown to all four evaluators for a given patient, the relative ranking

between the AI and the surgeons should remain valid (all four faced an identical handicap on the same subset of patients), but the absolute sensitivity and specificity figures in Table 3 may be underestimates of true performance for all evaluators. If, instead, site selection was performed independently for each evaluator rather than sharing a single selected image per patient, the paired comparison itself would be compromised, since evaluators would not necessarily have been assessing the same fracture.

Effective integration of this technology requires that AI-generated predictions be synthesized with a comprehensive review of the patient's clinical history, physical examination findings and operative details[25]. The cohort pools patients ranging from age 5 to 91 and fractures at five distinct anatomical sites treated with substantially different fixation techniques (intramedullary nails, plates, screws, elastic nails, cannulated screws, Kirschner wires); pediatric elastic nailing and adult intramedullary nailing, for example, are mechanically distinct scenarios that a single pooled accuracy figure may obscure. Subgroup-level breakdowns by anatomical site, age, and fixation technique were not available for the present analysis and are acknowledged as a limitation; we recommend them as a priority for any future validation cohort with sufficient subgroup sample sizes. Future research should also prioritize external, independent, prospective validation across diverse fracture patterns, fixation techniques, and imaging quality, and structured assessment of interobserver agreement among clinicians.

Despite its performance advantage in this cohort, the AI model shares a limitation common to deep neural networks: it lacks explainability. It produces a classification without indicating which radiographic features drove that output. This opacity is a recognized limitation across orthopaedic AI applications more broadly, where the inability to trace a model's reasoning back to identifiable anatomical or mechanical features undermines trust, complicates error analysis, and raises accountability concerns[31]. Explainable AI techniques would let trauma surgeons verify that the model's classifications rest on genuine features like screw purchase, fracture-line displacement and hardware alignment rather than incidental patterns in the training data[30]. Without such interpretability, even a model with a measurable performance advantage may face resistance to routine clinical integration. Future studies should adhere to standardized reporting frameworks for AI-based clinical research, such as the CONSORT-AI extension and the CLAIM framework, and should deposit datasets and analysis code in a public repository to support reproducibility and cross-study comparability[32,33].

Conclusion

The AI model outperformed all three participating trauma surgeons on every diagnostic metric evaluated, with statistically significant superiority over two of three surgeons; its advantage over the most experienced surgeon did not reach statistical significance in this cohort. Given the moderate class imbalance in the failure outcome, class-wise performance is a more informative summary than aggregate accuracy alone, and PPV and NPV must be interpreted in light of the 22.3% failure prevalence observed here. These results support a possible role for the model as a clinical decision-support adjunct rather than a clinician replacement, but this conclusion remains provisional pending external prospective validation on an independently sourced cohort. Subgroup analyses across anatomical site, age, and

fixation technique, and separation of mechanically- from biologically-driven failure in the primary endpoint, were not possible in the present analysis and are acknowledged as limitations rather than resolved questions; both should be prioritized in future work. Public deposition of the underlying dataset and analysis code is also still required. Establishing standardized reporting practices, including adherence to frameworks such as CONSORT-AI, will be important for building the transparency and explainability required for genuine clinician trust as this technology moves toward clinical integration.

Declarations

Ethics Statement

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and applicable institutional and national regulations governing research involving human participants. The study protocol was reviewed and approved by the respective Institutional Review Boards at all participating centres: B Lal Hospital and Research Centre, Sanchore (IRB approval no. BL/2024/092), Rahmania Medical Center, Motihari (IRB approval no. RMC/030624006), and Medi Plus Multi-Speciality Hospital, Sanchore (IRB approval no. MP/202406/054). Written informed consent was obtained from all participants before enrolment.

Consent to Participate

Written informed consent to participate in the study was obtained from all adult participants. For participants younger than 18 years, written informed consent was obtained from their parent or legal guardian, and age appropriate assent was obtained from the participant. Participants and their legal guardians were informed about the purpose of the study, the use of clinical and radiographic data, the voluntary nature of participation, and their right to withdraw from the study without affecting their clinical care.

Funding

This research did not receive funding.

Author Contribution

G.V. and D.S. contributed to the study conception and design, performed the clinical data collection, and wrote the main manuscript text. S.S. supervised the study and served as an independent reviewer, with no involvement in the surgical procedures. All authors contributed to the interpretation of the data, critically reviewed the manuscript, and approved the final submitted version.

Data Availability

The data are available upon request with the corresponding author.

References

1. Chawla, H. K., Rehncy, J. S., Bhatti, J., Sharma, M. & Bakshi, A. S. Management of Implant Failure in Long Bones and Its Functional Outcomes. Cureus. (2026). https://doi.org/10.7759/cureus.105877
2. Stuani, R. et al. Performance of Artificial Intelligence Models in Radiographic Image Analysis for Predicting Hip and Knee Prosthesis Failure: A Systematic Review. *Bioengineering* **13** (1), 122 (2026).
3. Dijkstra, H. et al. Systematic review of machine-learning models in orthopaedic trauma. *Bone Joint Open.* **5** (1), 9–19 (2024).
4. Lisacek-Kiosoglous, A. et al. Artificial intelligence in orthopaedic surgery. *Bone Joint Res.* **12** (7), 447–454 (2023).
5. Oosterhoff, J. H. F. et al. Feasibility of Machine Learning and Logistic Regression Algorithms to Predict Outcome in Orthopaedic Trauma Surgery. *J. Bone Joint Surg.* **104** (6), 544–551 (2021).
6. Marmor, M. T. et al. *Artificial intelligence: international perspectives on critical issues* 8 (OTA International, 2025).
7. Nijhuis, K. D. O. et al. AI for detection, classification and prediction of loss of alignment of distal radius fractures; a systematic review. *Eur. J. Trauma Emerg. Surg.* **50** (6), 2819–2831 (2024).
8. Olczak, J. et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop.* **88** (6), 581–586 (2017).
9. Kuo, R. et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. *Radiology* **304** (1), 50–62 (2022).
10. Lex, J. R. et al. Artificial Intelligence for Hip Fracture Detection and Outcome Prediction. *JAMA Netw. Open.*, **6**(3). (2023).
11. Glinkowski, W., Spalińska, A., Wołk, A. & Wołk, K. Artificial Intelligence in Orthopaedics: Clinical Performance, Limitations, and Translational Readiness—A Review. *J. Clin. Med.* **15** (5), 1751 (2026).
12. Dipnall, J. F. et al. Predicting fracture outcomes from clinical registry data using artificial intelligence supplemented models study protocol. PLoS ONE, 16(9). (2021).
13. Salimi, M., Parry, J. A., Shahrokhi, R. & Mosalamiaghili, S. Application of artificial intelligence in trauma orthopedics: Limitation and prospects. *World J. Clin. Cases*. **11** (18), 4231–4240 (2023).
14. Olczak, J. Fracture evaluation and prediction of outcome after a fracture using artificial neural networks. (2024).
15. Yapıcı, F. Who is the surgeon now: Human hands or machine minds? *J. Clin. Med.* **15** (6), 2165 (2026).
16. Oosterhoff, J. H. F. & Doornberg, J. N. Artificial intelligence in orthopaedics: false hope or not? *EFORT Open. Reviews*. **5** (10), 593–603 (2020).

17. Kim, S. & Lee, W. Does McNemar's test compare the sensitivities and specificities of two diagnostic tests? *Stat. Methods Med. Res.* **26** (1), 142–154. https://doi.org/10.1177/0962280214541852 (2017).
18. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochemia Med.* **22** (3), 276–282 (2012).
19. Jiang, J. et al. Noise-robustness test for ultrasound breast nodule neural network models as medical devices. *Front. Oncol.* **13**, 1177225 (2023).
20. Emanuel, E. J. *What Makes Clin. Res. Ethical? JAMA*, **283**(20), 2701. (2000).
21. Kalbas, Y. et al. Developments in the understanding of staging a major fracture in polytrauma. *Eur. J. Trauma Emerg. Surg.* **50** (3), 657–669 (2023).
22. Henderson, C. E. et al. Healing Complications Are Common After Locked Plating for Distal Femur Fractures. *Clin. Orthop. Relat. Res.* **469** (6), 1757–1765 (2011).
23. Ounasser, N., Rhanoui, M., Mikram, M. & Asri, B. E. A Systematic Review on Artificial Intelligence in Orthopedic Surgery. *Revue d'Intelligence Artificielle*. **38** (4), 1143–1157 (2024).
24. Zura, R., Mehta, S., Rocca, G. J. D. & Steen, R. G. Biological Risk Factors for Nonunion of Bone Fracture. *JBJS Reviews*, **4**(1). (2016).
25. Vaziri, R., Nazari, I., Dorri, F. & Dorri, F. Generative Artificial Intelligence for Trauma and Fracture Surgery. *Infosci. Trends*. **2** (5), 1–10 (2025).
26. Prijs, J. et al. Artificial intelligence and computer vision in orthopaedic trauma. *Bone Joint J.* **8**, 911–914 (2022).
27. Tian, C. et al. Artificial intelligence in orthopaedic trauma. *EngMedicine* **1** (2), 100020 (2024).
28. Wurm, L. M., Ertel, W. & Laue, D. Artificial intelligence in orthopedic trauma surgery: a scoping review. *Arch. Orthop. Trauma Surg.*, **146**(1). (2026).
29. Oettl, F. C., Zsidai, B., Oeding, J. F. & Samuelsson, K. Artificial Intelligence and Musculoskeletal Surgical Applications. *HSS J.* **21** (3), 267–273 (2025).
30. Oettl, F. C., Oeding, J. F. & Samuelsson, K. Explainable artificial intelligence in orthopedic surgery. *J. Experimental Orthop.*, **11**(3). (2024).
31. Zsidai, B. et al. A practical guide to the implementation of AI in orthopaedic research – part 1. *J. Experimental Orthop.* **10** (1), 117 (2023).
32. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* **370**, m3164. https://doi.org/10.1136/bmj.m3164 (2020).
33. Cerdá-Alberich, L. et al. MAIC–10 brief quality checklist for publications using artificial intelligence and medical images. *Insights into Imaging*. **14** (1), 11 (2023).