

Artificial Intelligence-Enabled Bone Fracture Assessment: A Comprehensive Review of Imaging and Sensing Technologies, Machine Learning Methods, and Clinical Integration

Auti Om Jitendra, Prof. Abhishek Popat Bangar, Prof. S. B. Bhosale, Dr. A. A. Khatr

Dept. of Computer Engineering
Jaihind College of Engineering, Pune, India

Abstract — Bone fractures are among the most common musculoskeletal injuries, and their outcome depends on accurate classification, timely localisation and early recognition of failed healing. Manual radiograph interpretation is subjective and time-consuming, and plain X-ray is slow to reveal healing failure (nonunion), which affects up to 12 % of tibial fractures and can take more than six months to diagnose [1]. Artificial intelligence (AI), combined with medical imaging and novel sensing, is increasingly proposed as a route to fast, consistent and early fracture assessment. This paper presents a focused narrative review of four verified peer-reviewed IEEE Access studies (2022–2025) covering contactless antenna-based sensing of implant deflection and fracture stiffness, X-ray fracture-type classification with transfer learning and ensemble classifiers, two-stage CT-based cervical spine fracture classification and vertebra detection, and attention-enhanced YOLOv8 detection of paediatric wrist fractures. The review examines three connected themes: imaging and sensing technologies (X-ray, computed tomography and near-field electromagnetic sensing), learning methods (CNN transfer learning, hybrid CNN–classical ensembles and object detectors), and clinical integration with explainability. The literature shows that deep models can classify fracture types with about 95–98 % accuracy, that YOLOv8 variants reach about 93.5 % mAP50 for vertebra detection and 65.8 % mAP50 over all wrist-trauma categories, and that a miniaturised Vivaldi antenna (36 × 30 mm) can track sub-millimetre implant displacement. However, most studies use single-source datasets, small test sets, random data splits and offline evaluation, and few report clinical validation. The review contributes a comparative analysis of representative studies, a proposed conceptual architecture linking acquisition, AI inference, decision support and clinical reference validation, and a prioritised agenda of research gaps and future directions.

Keywords— Bone Fracture, Deep Learning, YOLOv8, Transfer Learning, Medical Imaging, Fracture Healing Sensing.

I. INTRODUCTION

Fractures result from trauma such as falls, road accidents and sports injuries, or from bone-weakening conditions such as osteoporosis [2]. Treatment depends on correct classification of the fracture type, localisation of the affected bone, and monitoring of healing after fixation.

Traditional diagnosis relies on manual reading of radiographs, which is time-consuming and prone to human error, and inter-observer variability makes labelling itself subjective [2], [3]. Cervical spine fractures are a medical emergency that can cause permanent paralysis or death if missed [3]. After surgery, healing failure is hard to detect early: biplanar X-ray is qualitative and subjective, and mean nonunion diagnosis time exceeds six months, whereas fracture stiffness can predict healing outcome during the first 30 days [1].

AI is now applied across the fracture care pathway, from classification and detection on images to biomechanical monitoring of healing. For fracture assessment specifically, three developments converge: imaging and sensing modalities that capture the injury or the implant, ML and DL models that map signals to diagnoses, and explainable, workflow-integrated deployment that moves models into radiology practice [2]–[4]. This review examines these three strands together, asks how mature the evidence is for clinical deployment, and proposes a conceptual architecture that addresses the weaknesses identified.

II. REVIEW METHODOLOGY

This review is based on four peer-reviewed primary studies published in IEEE Access between 2022 and 2025 and supplied for this work [1]–[4]. Studies were included because they reported fracture-specific imaging, sensing or prediction with a

stated method and quantitative outcome. Supporting details such as dataset origins and baseline methods were taken from works cited inside those papers [5]–[7]. Every reference is traceable to a publisher record with a DOI or volume and page numbers. Direct logged querying of Scopus and Web of Science was not performed, so the review is narrative rather than systematic, and study counts beyond the four primary papers are not reported.

III. IMAGING AND SENSING TECHNOLOGIES FOR FRACTURE ASSESSMENT

1. Near-Field Electromagnetic Sensing of Implants

Metal fixation hardware (rods, plates and screws) stabilises most long-bone fractures, and the bone–implant pair behaves as a composite whose stiffness changes as the callus stiffens. A coiled coaxial dipole antenna previously allowed noninvasive detection of implant displacement through shifts in antenna resonant frequency, but its large size and highly nonlinear response to the antenna–implant distance limited clinical use [1], [5]. Wolynski et al. developed two planar Vivaldi antennas on FR-4 PCB for operation near 1.5 GHz: a “standard” design (138×100 mm) and a “miniaturised” design (36×30 mm), an approximately 13-fold area reduction [1]. In benchtop tests, resonant frequency of the miniaturised antenna rose by 227.2 MHz (parallel) and 98.5 MHz (perpendicular) over 0–10 mm of stainless-steel plate displacement, and by 177.7 and 58.1 MHz for titanium, with sensitivity largest below 1 mm and decreasing nonlinearly with distance [1]. Because shifts were similar whether the plate or the antenna was moved (average difference 5.2 % for distances of 0.5 mm or more, never above 10 %), a known antenna displacement can calibrate the diagnostic data and convert frequency shift into implant deflection, the inverse of bending stiffness [1].

2. Radiography (X-Ray)

X-ray is the most widely used modality because of cost-effectiveness and availability [4]. Torne et al. used 1,129 grey-scale X-ray images ($256 \times 256 \times 3$) from a public Kaggle collection covering ten fracture types (avulsion, comminuted, fracture-dislocation, greenstick, hairline, impacted, longitudinal, oblique, pathological and spiral), with 989 images for training and 140 for testing [2]. Chien et al. worked with GRAZPEDWRI-DX, a public set of 20,327 paediatric wrist trauma radiographs with 74,459 image labels collected at one hospital between 2008 and 2018 [4], [7]. Both datasets are imbalanced: in GRAZPEDWRI-DX, bone anomaly and soft tissue make up only 0.41 % and 0.68 % of labelled objects [4].

3. Computed Tomography (CT)

CT is a reliable method for identifying cervical spine fractures [3]. Yaseen et al. used the RSNA 2022 Cervical Spine Fracture Detection dataset of 2,019 CT scans, each with 100–800 slices. DICOM images were converted to PNG with a window centre of 500 and width of 2000, slices with bounding-box annotations were labelled positive and the rest negative, and about 7,217 positive images were balanced by under-sampling the negative class [3]. Bounding boxes for vertebrae C1–C7 were derived from segmentation labels, yielding 9,170 images for detection [3].

IV. MACHINE LEARNING AND DEEP LEARNING METHODS

1. Transfer Learning Versus Hybrid Ensembles

Torne et al. compared four pipelines: VGG-16 with custom dense layers, and pre-trained feature extractors feeding classical classifiers (VGG-16 with random forest, ResNet-50 with a linear-kernel SVM, EfficientNetB0 with XGBoost). VGG-16 and VGG-16 with random forest both reached 0.95 accuracy, ResNet-50 with SVM reached 0.93, and EfficientNetB0 with XGBoost only 0.41, with sensitivity for impacted fractures as low as 0.06 [2]. The hypothesis that ensembles would outperform individual networks was therefore not supported: the plain fine-tuned VGG-16 matched the hybrid. VGG-16 is also the most expensive (about 40.1 GFLOPs and 14.7 M parameters versus 10.1 GFLOPs for ResNet-50 and 0.8 GFLOPs for EfficientNetB0) [2]. The poor EfficientNetB0 result was attributed by the authors to weak synergy between its features and gradient boosting on small greyscale images, an explanation that was not tested by ablation.

2. Classification with Explainability

Yaseen et al. proposed a sequential CNN (Conv2d, ReLU, MaxPool2d, flatten and three linear layers) to classify CT slices as fractured or not fractured. On an 80:20 split it achieved precision 98.1 %, recall 97.5 %, F1 97.8 %, accuracy 97.8 %, Cohen’s kappa 96.6 % and specificity 99.0 %, ahead of ResNet50 (94.9 % accuracy) and DenseNet121 (94.2 %) [3]. An ablation showed accuracy rising from 92.2 % to 94.6 % to 97.8 % as fully connected layers increased from one to three [3]. Grad-CAM heat maps highlighted the regions driving each decision, supporting verification by clinicians [3].

3. Object Detection

Detection answers where the fracture or bone is. In the second stage of Yaseen et al., YOLOv5l and four YOLOv8 sizes were trained at 512×512 with SGD (learning rate 0.01, 200 epochs,

batch 16). YOLOv8l was best, with precision 90.0 %, recall 89.0 %, mAP50 93.5 % and mAP50-95 87.2 % for vertebra detection [3]. Chien et al. inserted attention modules into the YOLOv8 neck at four positions and compared ResCBAM, shuffle attention, efficient channel attention and ResGAM on GRAZPEDWRI-DX. At 1024 input size and large model size, YOLOv8+ResCBAM raised mAP50 from 63.6 % to 65.8 % and recall from 67.9 % to 71.6 %, with inference time rising from 7.7 ms to 8.7 ms; ResGAM reached 65.0 % but needed 18.1 ms [4]. For the single “fracture” label, all variants exceeded 95 % mAP50 (95.7 % for ECA) [4]. A medium ResCBAM model at 5.7 ms retained 64.3 % mAP50, a practical accuracy–speed compromise [4].

4. Validation Practice

Validation designs vary and are often under-reported. Yaseen et al. and Chien et al. divide data randomly (80:20 and 70:20:10 respectively) [3], [4]. Neither reports patient-level grouping, and a CT scan contributes hundreds of correlated slices, so slices from one patient may appear in both training and test sets, which can inflate accuracy. Torne et al. report a test set of only 140 images (9–19 per class), so per-class scores such as the 1.00 values in [2] are statistically fragile. Grouped validation and external test sets remain uncommon in fracture AI studies.

V. CLINICAL INTEGRATION AND DEPLOYMENT

Deployment studies differ mainly in how the model reaches the clinician. Torne et al. describe linking classifiers to hospital imaging systems so that a radiologist uploading an X-ray receives an immediate fracture-type prediction, while noting the compute demands of VGG-16, the need for HIPAA-style privacy compliance and continuous retraining [2]. Chien et al. plan web and mobile applications and a staged clinical validation, from a 100-patient pilot to 1,000 patients across at least three institutions and regulatory approval [4].

Yaseen et al. position the two-stage system as support for radiologists in emergency settings, with Grad-CAM for transparency [3]. The antenna approach is the only reviewed method that needs hardware at the patient: a thin, small Vivaldi sensor mounted on an actuator to perform calibration during follow-up visits, with future arrays for higher resolution [1].

None of the studies reported prospective operation inside a hospital workflow, and the antenna work remains benchtop-validated, with translational studies still recommended [1].

VI. COMPARATIVE ANALYSIS

Table I summarises representative primary studies. Accuracy values are not directly comparable because targets, datasets and validation schemes differ.

Table 1: Comparison of Representative AI-Based Fracture Assessment Studies

Author, year	Objective	Sensing / method	Major reported findings	Limitations
Wolynski et al., 2022 [1]	Sense implant deflection and stiffness	Vivaldi antenna (36 × 30 mm); resonant frequency shift	Up to 227.2 MHz shift over 10 mm; ~13-fold smaller than standard design	Benchtop plates only; no tissue or in vivo test
Yaseen et al., 2024 [3]	Detect fracture and locate C1–C7	CT; custom CNN + Grad-CAM; YOLOv5/YOLOv8	97.8 % accuracy; YOLOv8l mAP50 93.5 %	Random slice split; single public dataset
Torne et al., 2025 [2]	Classify 10 fracture types	X-ray; VGG-16, VGG-16 + RF, ResNet-50 + SVM, EfficientNetB0 + XGBoost	0.95 accuracy (VGG-16 and VGG-16 + RF); 0.41 for EfficientNetB0 + XGBoost	140 test images; class imbalance
Chien et al., 2025 [4]	Detect paediatric wrist fractures	X-ray; YOLOv8 + ResCBAM, SA, ECA, ResGAM	mAP50 65.8 % (ResCBAM, 1024); 95.7 % for fracture label	Single institution; rare classes weak

VII. PROPOSED CONCEPTUAL ARCHITECTURE

Fig. 1 presents a proposed conceptual architecture derived from the reviewed systems. A fracture is imaged at presentation, or a fixed fracture is followed up (layer 1). Complementary modalities capture X-ray, CT and, for implanted limbs, near-field antenna signals (layer 2), because no single modality covers diagnosis, localisation and healing monitoring [1]–[4]. An acquisition step converts DICOM data, applies CT windowing, resizing and augmentation, and performs antenna calibration using a known displacement (layer 3) [1], [3], [4]. AI models then perform fracture-presence and fracture-type classification, YOLO-based localisation and stiffness estimation from calibrated frequency shift, each with a confidence score and Grad-CAM explanation (layer 4) [1]–[4].

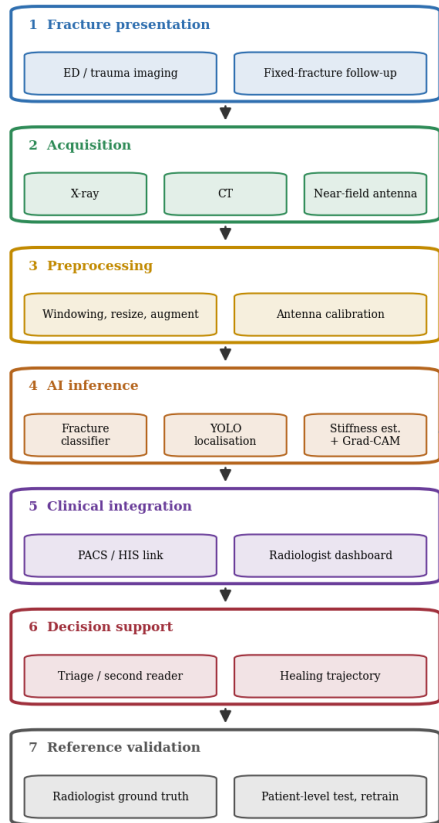


Fig.1. Proposed conceptual architecture for AI-enabled bone fracture assessment from presentation to clinical validation (conceptual, not experimentally validated)

Results reach a radiologist dashboard through the hospital imaging system (layer 5) [2], and decision rules support triage, second-reader alerts and healing-trajectory tracking, with the clinician retaining the final decision (layer 6). A sampled subset is checked against radiologist ground truth with patient-level held-out testing (layer 7), and models are retrained only after independent re-testing, which addresses the validation gaps identified in the literature [2]–[4].

VIII. RESEARCH GAPS AND FUTURE DIRECTIONS

Gaps supported by the reviewed evidence. (i) Single-source, imbalanced datasets: each imaging study uses one public dataset, and rare classes perform poorly [2], [4]. (ii) Small or non-independent test sets: 140 test images in [2] and random slice-level splits in [3], [4]. (iii) Offline evaluation: no study reports prospective hospital deployment, and the antenna work

is benchtop only [1]. (iv) Limited cross-institution generalisability, acknowledged by the authors of [4]. (v) Untested explanations: reasons for the failure of EfficientNetB0 with XGBoost and for the weak GAM gain were hypothesised but not ablated [2], [4]. (vi) Computational cost: the best classifier needs about 40 GFLOPs [2], and some attention modules double inference time [4].

Directions requiring investigation. Priorities include multi-centre datasets with radiologist-verified labels; patient-level grouped validation and external test sets; class-imbalance remedies for rare fracture types; lightweight models for resource-limited settings; transformer-based or multi-head attention detectors, as suggested in [4]; translational studies of calibrated antenna sensing in animal and clinical models, and antenna arrays for resolution [1]; and fusion of imaging-based diagnosis with biomechanical healing metrics. Data privacy and regulatory approval must be addressed alongside accuracy [2], [4].

IX. CONCLUSION

AI-enabled fracture assessment has progressed from manual radiograph reading to automated classification, vertebra-level localisation with explainable outputs, attention-enhanced detection and compact sensors for implant-based healing monitoring. Reported performance is high for specific tasks, such as fracture-type classification at about 95 % accuracy and CT slice classification at 97.8 %, but the evidence base is dominated by single-dataset, offline studies with limited independent validation. Simple fine-tuned networks frequently match more complex hybrids, and clinical reference standards remain essential. Reliable deployment will require larger and more diverse datasets, patient-level validation, prospective clinical trials, multi-modal fusion and continuous reference-based model maintenance, as reflected in the proposed conceptual architecture.

REFERENCES

1. J. G. Wolynski, M. M. Ilić, B. M. Notaroš, K. M. Labus, C. M. Puttlitz, and K. C. McGilvray, "Vivaldi antennas for contactless sensing of implant deflections and stiffness for orthopaedic applications," *IEEE Access*, vol. 10, pp. 1151–1161, 2022, doi: 10.1109/ACCESS.2021.3137718.
2. S. Torne, D. K. Shetty, K. Makkithaya, P. Hegde, M. Sudhi, P. K. Pullela, T. Eniyan T, R. Kamath, S. Salu, P. Bhat, S. Girisha, and P. S. Priya, "VGG-16, VGG-16 with random forest, ResNet50 with SVM, and EfficientNetB0 with XGBoost-enhancing bone fracture classification in X-ray

- using deep learning models,” IEEE Access, vol. 13, pp. 25568–25577, 2025, doi: 10.1109/ACCESS.2025.3534818.
3. M. Yaseen, M. Ali, S. Ali, A. Hussain, M.-I. Joo, and H.-C. Kim, “Cervical spine fracture detection and classification using two-stage deep learning methodology,” IEEE Access, vol. 12, pp. 72131–72142, 2024, doi: 10.1109/ACCESS.2024.3398061.
 4. C.-T. Chien, R.-Y. Ju, K.-Y. Chou, E. Xieerke, and J.-S. Chiang, “YOLOv8-AM: YOLOv8 based on effective attention mechanisms for pediatric wrist fracture detection,” IEEE Access, vol. 13, pp. 52461–52477, 2025, doi: 10.1109/ACCESS.2025.3549839.
 5. K. M. Labus, B. M. Notaros, M. M. Ilic, C. J. Sutherland, A. Holcomb, and C. M. Puttlitz, “A coaxial dipole antenna for passively sensing object displacement and deflection for orthopaedic applications,” IEEE Access, vol. 6, pp. 68184–68194, 2018.
 6. R.-Y. Ju and W. Cai, “Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm,” Sci. Rep., vol. 13, no. 1, Art. no. 20077, Nov. 2023.
 7. E. Nagy, M. Janisch, F. Hrzić, E. Sorantin, and S. Tschauner, “A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning,” Sci. Data, vol. 9, no. 1, Art. no. 222, May 2022.