e-ISSN: 0976-822X, p-ISSN:2961-6042

Available online on http://www.ijcpr.com/

International Journal of Current Pharmaceutical Review and Research 2026; 18(8); 880-892

Original Research Article

Machine Learning for Predicting Fracture Healing and Non-Union: A Narrative Review

Narendra B. S.[1], Prajwal J.[2], Pramod G.[3]

[1,2,3]Department of Orthopaedics, Chikkaballapur Institute of Medical Sciences, Rajiv Gandhi University of Health Sciences, Chikkaballapur, Karnataka, India

Received: 01-05-2026 / Revised: 15-06-2026 / Accepted: 21-07-2026

Corresponding author: Dr. Pramod G

Conflict of interest: Nil

Abstract

Background: Fracture healing assessment remains subjective and delayed, relying on serial radiographs interpreted with semi-quantitative scoring systems (RUST, RUSH, RUSHU) that carry substantial observer variability. Machine learning (ML) offers the prospect of earlier, objective, and individualized prognostication of union, delayed union, and non-union.

Methods: We performed a narrative evidence-synthesis review prepared with reference to the SANRA checklist. PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, and preprint servers were searched (January 2015–June 2026); thirty representative publications on ML applications were synthesized across four domains: image-based deep learning, mechanobiology-informed finite-element/ML hybrid models, wearable-sensor biomechanics, and registry-based mortality/complication prediction.

Results: Convolutional networks dominate image-based tasks, with union-prediction accuracies of 91–93.6% in scaphoid series and detection AUCs up to 0.97; ML surrogates reproduce finite-element healing simulations at a fraction of the computational cost; wearable insole data predicted lower-extremity healing with approximately 76% accuracy (n = 25); and ML models predicted post-hip-fracture mortality with AUCs of 0.68–0.89 (largest cohort 74,396 cases), with gradient-boosting methods frequently outperforming logistic regression. SHAP-based explanations consistently revealed biologically plausible predictors. Nearly all studies, however, are retrospective, single-center, and internally validated, with heterogeneous non-union definitions and infrequent external validation.

Conclusions: ML-based fracture healing prediction is promising but remains at a proof-of-concept stage. Multi-center prospective datasets, standardized outcome definitions with TRIPOD/CLAIM-concordant reporting, external validation, and clinician-in-the-loop trials are prerequisites for clinical adoption; no current model is ready for standalone clinical decision-making.

Keywords: Machine Learning; Deep Learning; Fracture Healing; Bone Union; Non-Union; Radiographic Union Score; Artificial Intelligence; Orthopedic Trauma; Explainable Artificial Intelligence; Predictive Modeling.

DOI: 10.25258/ijcpr.18.8.136

Introduction

Bone fractures are among the most common musculoskeletal injuries worldwide. Although the majority heal uneventfully with standard fixation or conservative care, non-union complicates approximately 10% of fractures overall, with substantially higher rates after open and high-energy injuries and in the presence of patient-related risk factors such as smoking, elevated body mass index, and alcohol use [1]. Delayed union and non-union result in prolonged disability, repeated surgical intervention, and substantial economic burden [1,2]. Timely identification of fractures at risk of impaired healing is therefore critical to guide early intervention, whether through modification of mechanical loading, biological augmentation, or revision surgery. Traditional approaches to assessing fracture healing rely heavily on serial plain radiography interpreted using semi-quantitative scoring systems: the Radiographic Union Score for Tibial fractures (RUST) [3], validated for reliability and validity [4]; its hip analogue, the Radiographic Union Score for Hip fractures (RUSH) [5,6]; and the humeral analogue RUSHU, which predicts humeral shaft non-union [7]. These are complemented by clinical criteria including pain on weight-bearing and tenderness at the fracture site [3–7]. Although clinically useful, these tools suffer from considerable inter- and intra-observer variability and generally detect impaired healing only after a substantial delay, limiting opportunities for early corrective intervention [4–8]. Over the past decade, artificial intelligence (AI) and its principal subfield, machine learning (ML), have been increasingly applied across orthopedic trauma care, spanning automated fracture detection, injury classification, surgical planning, and outcome prediction [9–11]. Within this broader landscape, a growing and methodologically diverse body of work has specifically targeted the prediction of fracture healing trajectories, non-union risk, and associated complications, using data ranging from radiographic images and micro-computed tomography (micro-CT) scans [12] to wearable-sensor gait metrics [13] and structured clinical registries [14]. Unlike fracture detection — now a comparatively mature application with radiologist-level performance reported in several studies [15,16] — healing prediction poses distinct challenges, because it requires modeling a dynamic, weeks-to-months-long biological process rather than interpreting a single static image.

This review aims to (i) summarize the biological basis of fracture healing relevant to model design, (ii) synthesize the major ML approaches applied to fracture healing and related outcome prediction, (iii) present a comparative overview of reported model performance across application domains, (iv) discuss the role of model interpretability in clinical adoption, and (v) identify the principal methodological gaps and future research priorities in this rapidly evolving field. The organizational taxonomy of the approaches reviewed is summarized in Figure 2.

Biological Basis of Fracture Healing: An understanding of the underlying biology of fracture repair is essential context for interpreting the features, inputs, and time horizons used by predictive models. Fracture healing is classically described as a four-phase, partially overlapping process: an initial inflammatory phase marked by hematoma formation and recruitment of immune cells; a soft-callus phase in which mesenchymal progenitor cells differentiate into chondroblasts and fibroblasts, forming a cartilaginous bridging matrix; a hard-callus phase in which this cartilaginous template undergoes endochondral ossification into woven bone; and a prolonged remodeling phase during which woven bone is progressively replaced by lamellar bone until the original cortical architecture is restored [2,17]. Immune–skeletal crosstalk (osteoimmunology) is now recognized as a central regulator of this sequence, linking the early inflammatory milieu to later reparative and remodeling events [2]. Mechanical stability at the fracture site, vascular supply, and local biological signaling — including growth factors such as transforming growth factor-β, vascular endothelial growth factor, and bone morphogenetic proteins — jointly govern whether healing proceeds along a normal trajectory or is diverted toward delayed union or non-union [2,17]. Mechanobiological theory further posits that the mechanical strain and interfragmentary movement experienced at the fracture gap regulate tissue differentiation, providing the conceptual foundation for finite-element (FE) models of fracture healing that simulate stress, strain, and tissue differentiation over time [17]. These physics-based models, although mechanistically detailed, are computationally expensive, which has motivated the development of ML surrogates capable of approximating FE outputs at a fraction of the computational cost [18,19].

Review Methodology: This narrative review was prepared with reference to the SANRA (Scale for the Assessment of Narrative Review Articles) checklist. A structured literature search of PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, and preprint servers (arXiv, medRxiv) was performed for publications dated January 2015 to 30 June 2026, using combinations of ("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural network") AND ("fracture healing" OR "bone union" OR "non-union" OR "delayed union" OR "fracture repair"), without restriction on anatomical site; a representative PubMed search string is provided in Appendix A.

Eligible publications were peer-reviewed original studies and systematic reviews in English reporting ML or deep-learning models for fracture healing, union/non-union classification, healing-stage assessment, or fracture-related mortality and complication prediction. Foundational biomechanical and biological literature predating 2015 was included where conceptually necessary. Titles and abstracts were screened by two authors, and thirty representative publications addressing ML applications were purposively selected to span image-based deep learning, mechanobiology-informed hybrid modeling, wearable-sensor approaches, and registry-based prediction of fracture-related mortality and complications; the selection flow is summarized in Figure 1, and additional foundational references describing scoring-system development and fracture biology were cited where relevant. Selection was purposive rather than exhaustive, and the resulting synthesis should be interpreted as a representation of the current evidence landscape.This review does not report a formal risk-of-bias assessment (e.g., PROBAST) or meta-analytic pooling of effect sizes and should therefore be interpreted as a qualitative evidence synthesis rather than a quantitative systematic review.

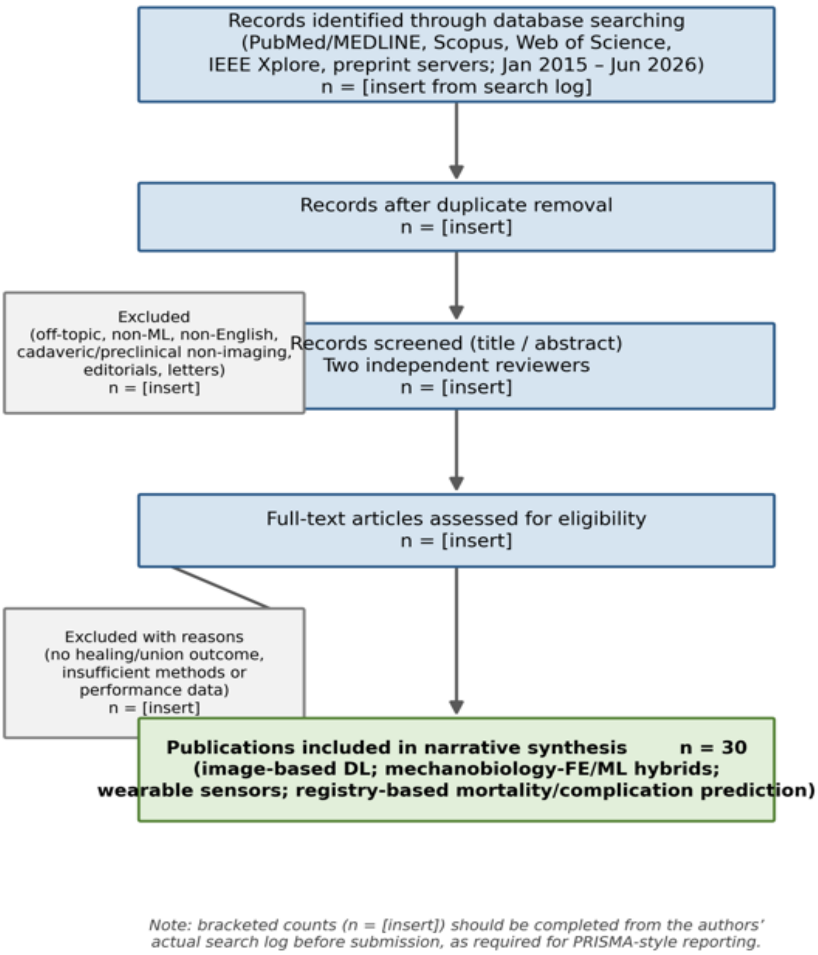

Figure 1: Literature identification and selection flow. Bracketed counts should be completed from the authors' search log prior to submission.

Machine Learning Approaches for Fracture Healing Prediction

Radiographic and Imaging-Based Deep Learning Models: The largest and most mature body of work applies convolutional neural networks (CNNs) and related deep learning architectures to plain radiographs, computed tomography (CT), and micro-CT for fracture detection, classification, and, increasingly, healing-stage assessment. Early CNN-based systems demonstrated the feasibility of automated fracture detection on wrist, scaphoid, hip, and whole-body trauma imaging, achieving sensitivity and specificity comparable to or exceeding that of practicing radiologists in several series [15,16,20,21]. A recent systematic review of deep learning in scaphoid care found that transfer-learning architectures such as ResNet, DenseNet, and EfficientNet consistently outperformed custom-built CNNs, particularly when trained on multicenter datasets, and that multi-view fusion of anteroposterior and lateral radiographs improved recall by approximately 12 percentage points relative to single-view analysis [22]. Beyond generic fracture detection, deep neural networks have also been applied to more specialized diagnostic distinctions, such as differentiating atypical femoral fractures from typical femoral shaft fractures on plain radiographs — a distinction with direct implications for identifying bisphosphonate-associated healing complications [23]. Building on fracture-detection capabilities, several groups have extended deep learning to the prediction of union versus non-union. In the largest series identified, a TensorFlow-based deep learning algorithm applied to pre- and post-operative radiographs of 346 scaphoid non-union patients predicted union following revision surgery with approximately 93.6% accuracy, notably using imaging data alone, without patient-level clinical covariates [24]. In the same evidence base, a custom CNN achieved approximately 93.6% accuracy in predicting post-surgical union, while a YOLOv5–ResNet-50 pipeline classified outcomes into union, non-union, and osteonecrosis categories with 91% accuracy and an area under the curve (AUC) of 0.96; across detection studies, reported accuracies ranged from 81% to 96% with AUCs up to 0.97 [22].

At the tissue level, a segmentation-based deep network (VM-TE-UNet) applied to micro-CT imaging of rodent fracture callus (2,448 images) achieved a Dice similarity coefficient of 0.809 for callus and non-union region segmentation (95th-percentile Hausdorff distance 13.1), improving on a baseline architecture, and identified quantitative bone morphometric parameters — including bone surface-to-tissue-volume ratios of mineralized cartilage and osteogenic tissue — that differed significantly between healing and non-union trajectories during the inflammatory and repair phases [12]. In that study, the model screened healing failure as early as 14 days after fracture in rats. Such findings illustrate how image-derived quantitative biomarkers can supplement, or eventually replace, qualitative radiographic scoring, although the rodent healing biology and timeline differ substantially from human fracture repair and warrant caution when extrapolating.

Non-Union and Delayed-Union Risk Prediction Models: A parallel and clinically important stream of research targets the prediction of non-union or delayed union using a combination of radiographic scoring, clinical risk factors, and, increasingly, ML classifiers.

Established radiographic scoring systems — RUST for tibial fractures [3,4], RUSH for hip fractures [5,6], and RUSHU for humeral shaft fractures [7] — provide semi-quantitative healing scores derived from serial follow-up radiographs and have demonstrated reasonable inter- and intra-observer reliability, forming a natural feature input for downstream predictive models [8]. In humeral shaft fractures managed non-operatively, a RUSHU score of eight or less at six weeks was associated with a 4.3-fold increased risk of non-union (relative risk 4.3; 95% CI 2.7–6.9), and applying the score improved non-union prediction from a 41% baseline risk to 66% at six weeks and 75% at twelve weeks; a cut-off of seven or less yielded 82% specificity at 63% sensitivity [8].

A recent systematic review and meta-analysis of prediction models for delayed union and non-union highlighted that, although numerous multivariable models have been developed, external validation remains infrequent and reported discrimination is highly variable across anatomical subgroups and validation tiers, underscoring the early stage of clinical maturity in this domain [26]. Clinical risk-factor analyses — several using logistic regression rather than more complex ML algorithms — have consistently identified high-energy trauma mechanism, open fracture, degree of cortical contact, postoperative infection, and use of staged external fixation as predictors of delayed union or non-union following intramedullary nailing of tibial shaft fractures, concordant with large epidemiologic analyses identifying open fracture, smoking, and obesity among the strongest non-union risk factors [1,26]. These logistic-regression-based studies, although not always described as "machine learning" in the strict sense, form an important methodological bridge toward the more complex ensemble and deep learning classifiers now being explored for the same prediction task.

Complementing score-based approaches, several authors have asked whether deep learning can predict non-union directly from radiographic appearance, reasoning that CNNs may detect sub-visual pixel-level changes in callus earlier than the human eye can appreciate bridging under RUST criteria [25]. Computed tomography, increasingly used as a reference standard for union assessment, represents a further promising — but still underexplored — modality for automated volumetric healing assessment.

Mechanobiology-Informed and Finite-Element–Machine Learning Hybrid Models: A distinctive and increasingly influential methodological direction couples mechanobiological FE simulation with ML surrogates, aiming to preserve the mechanistic fidelity of physics-based healing models while dramatically reducing computational cost. In one representative study, seven ML algorithms — support vector regression (SVR), random forest (RF), XGBoost, multilayer perceptron (MLP), CNN, recurrent neural network (RNN), and long short-term memory (LSTM) network — were trained on 648 simulated healing trajectories generated from a validated 21-day mechanoregulation-based FE model of femoral fracture healing in rodents, varying implant configuration, loading regime, and biological parameters [18]. The resulting ML surrogate generalized well to unseen parameter combinations, remained robust when trained on as little as 50% of the available data, and enabled Shapley-value-based interpretation showing that screw number increased central callus stiffness while excessive mechanical loading impaired outer callus formation — findings consistent with established mechanobiological theory [17,18].

A related computational modeling study of distal radius fracture healing integrated mechanoregulated cell differentiation, tissue formation, and angiogenesis into a three-dimensional FE model, generating 3,600 simulated cases across physiologically relevant loading conditions, fracture geometries, gap sizes, and healing times, then trained multiple ML algorithms on the simulation outputs to predict healing outcomes at different post-injury stages; a cubic support vector machine performed best during early-stage healing, whereas a trilayered artificial neural network outperformed alternatives during late-stage healing [19]. Model outputs also suggested that fracture geometry and gap size materially influenced the predicted healing trajectory — Smith fractures with medium gap sizes favored healing, whereas Colles fractures with large gap sizes led to delayed healing through excessive fibrous tissue formation [19].

These hybrid approaches illustrate a broader trend of using ML not merely as a black-box classifier but as a computationally efficient, explainable surrogate for mechanistic biological simulation, with potential application to real-time, patient-specific rehabilitation planning.

Wearable-Sensor and Biomechanical Data-Driven Models: A smaller but methodologically distinct body of work leverages continuous, objective biomechanical data captured via wearable sensors to predict fracture healing progression, addressing the subjectivity and intermittent nature of radiographic follow-up.

In a retrospective study of 25 patients with closed lower-extremity fractures, continuous underfoot loading data from a gait-monitoring insole were processed to extract step-level features, with decision-tree and LASSO-regression-based feature selection feeding a logistic regression classifier that predicted days-to-healing within a 30-day window, achieving approximately 76% mean accuracy, precision, recall, and F1-score under 10-fold and leave-one-out cross-validation [13]. Feature importance analysis highlighted the clinical relevance of medial underfoot loading distribution as a marker of healing progression, and the authors noted the approach's potential to enable earlier detection of complications such as asymptomatic delayed union [13].

Given the modest sample size and purely internal cross-validation, this work should be regarded as proof-of-concept-level evidence, with attention to the potential for optimistic performance when feature selection is performed within the cross-validation loop; external validation in larger cohorts is required. Nevertheless, it exemplifies a broader opportunity to integrate continuous, remote, and objective biomechanical monitoring into fracture healing surveillance, complementing rather than replacing periodic radiographic assessment.

Prediction of Mortality and Complications in Fracture Populations: Beyond direct prediction of bone union, a substantial and clinically influential body of ML research addresses prediction of mortality and major complications following fracture, particularly in hip fracture populations, where post-fracture mortality is a critical outcome for shared decision-making. Multiple independent groups have developed and internally validated ML models — including random forest, XGBoost, gradient boosting, generalized linear models, and naïve Bayes classifiers — for 30-day, 1-year, and in-hospital mortality following hip fracture, typically using preoperative demographic, biochemical, and comorbidity data [27–31]. Reported discriminative performance has generally ranged from an AUC of approximately 0.68 to 0.89 depending on cohort size, outcome horizon, and feature set, with gradient-boosting methods such as XGBoost frequently outperforming logistic regression and other classifiers in head-to-head comparisons [28,30]; a naïve Bayes classifier trained on sociodemographic and comorbidity data alone performed on par with more complex ensemble methods while remaining directly interpretable [31].

A large-scale application using the Dutch Hip Fracture Audit, encompassing 74,396 cases, extended this work toward shared decision-making, aiming to support individualized treatment discussions between clinicians and patients regarding operative versus non-operative management [14].

Although these models do not directly predict bone healing, they represent a mature and clinically impactful application of the same core ML methodologies within the broader fracture care pathway. Their validation experience offers instructive lessons for healing-prediction model development: most models report discrimination (AUC) but comparatively few report calibration or decision-curve analyses, and where external validation has been attempted, performance often degrades relative to internal estimates [10,14,26].

Broader Applications of Artificial Intelligence in Orthopedic Trauma: Several recent scoping and systematic reviews have mapped the wider landscape of AI applications in orthopedic trauma, situating fracture healing prediction within a broader ecosystem that includes automated fracture detection and classification, injury severity triage, surgical planning, and postoperative complication prediction [9–11,32–34].

A comprehensive review analyzing 217 studies published between 2015 and 2025 found that deep learning approaches (43.3%) and traditional machine learning methods (39.2%) together dominated the field, with fracture detection and classification representing the most common application category, followed by outcome prediction tasks; notably, only 14.5% of studies underwent external validation and just 3.2% reported prospective clinical validation [11]. A systematic review and meta-analysis of AI in trauma triage found that the best-performing models outperformed conventional triage tools for mortality, hospitalization, and critical-care admission prediction [32], while broader syntheses of AI in orthopedic surgery report high diagnostic accuracy and improving predictive performance across the surgical care pathway [33,34]. These broader reviews consistently emphasize that, while diagnostic and detection applications of AI have reached a relatively mature stage of validation, prediction tasks — including fracture healing prediction specifically — remain comparatively underdeveloped in terms of external validation and prospective clinical evaluation [9–11].

Comparative Summary of Representative Studies: Table 1 summarizes representative machine learning and deep learning approaches applied across the fracture healing prediction domains discussed above, illustrating the diversity of data modalities, algorithms, validation strategies, and reported performance metrics. Table 2 presents study-level characteristics of the principal healing- and outcome-prediction studies. NR = not reported; DRF = distal radius fracture.

Table 1: Representative machine learning approaches for fracture healing and related outcome prediction, with validation type and explainability

Domain	Representative studies (ref)	ML/DL method(s)	Data source / modality	Sample	Validation	Key reported performance	XAI
Mechanobiology–ML surrogate models	Razavi et al. 2025 [18]	SVR, RF, XGBoost, MLP, CNN, RNN, LSTM	Finite-element simulated healing trajectories (rodent femur)	648 simulated trajectories	Internal (unseen parameter combinations)	High agreement with FE ground truth; robust with as little as 50% training data	SHAP
Distal radius healing-stage prediction	Liu et al. 2023 [19]	Cubic SVM; trilayered ANN	3-D computational healing model outputs	3,600 simulated cases	Internal	Cubic SVM best at early stage; ANN best at late stage; gap size and geometry influential	NR
Wearable-sensor healing prediction	North et al. 2024 [13]	Decision tree, LASSO regression, logistic regression	Insole ground-reaction-force data	25 patients	10-fold and leave-one-out CV (internal)	≈76% accuracy, precision, recall, F1 for days-to-healing (30-day window)	Intrinsic (white-box models)
Micro-CT non-union prediction	Yu et al. 2025 [12]	VM-TE-UNet (deep segmentation) + early-diagnosis model	Rodent micro-CT callus images	2,448 images	Internal	Dice similarity coefficient 0.809; 95th Hausdorff distance 13.1; screening feasible at day 14	NR
Scaphoid union classification	Tümen et al. 2025 [24]; Jha et al. 2025 [22]	CNN, YOLOv5–ResNet-50, transfer learning (DenseNet, ResNet, EfficientNet)	Wrist/scaphoid radiographs	346 patients (non-union series); 14 studies in review	Internal	93.6% union-prediction accuracy; 3-class outcome accuracy 91% (AUC 0.96); detection accuracy 81–96%, AUC up to 0.97	Grad-CAM++ (detection studies)
Radiographic union scoring	Whelan et al. 2010 [3]; Frank et al. 2016 [6]; Oliver et al.	Radiographic scoring (RUST, RUSH,	Serial radiographs	Clinical cohorts	Observer reliability studies	Good inter-/intra-observer reliability; RUSHU ≤8 at 6	N/A

	2019 [7]; Suter et al. 2024 [8]	RUSHU) integrated with statistical/ ML models				weeks → RR 4.3 for non-union; used as ML input feature	
Hip-fracture mortality prediction	Mosfeldt et al. 2024 [27]; Kitcharanant et al. 2022 [28]; Li et al. 2021 [29]; Shuchami et al. 2026 [30]; Dijkstra et al. 2025 [14]; Wang 2025 [31]	Random forest, XGBoost, gradient boosting, logistic regression, naïve Bayes	Registry and administrative clinical data	Up to 74,396 cases (Dutch Hip Fracture Audit)	Internal validation; external validation infrequent	AUC 0.68–0.89 across cohorts; gradient boosting frequently outperforms logistic regression	SHAP / direct interpretation in some
General fracture detection (contextual; non-healing-specific)	Thian et al. 2019 [20]; Hendrix et al. 2021 [15]; Cheng et al. 2019 [16]; Lu et al. 2026 [21]; Zdolsek et al. 2021 [23]	CNN, object detection (YOLO-family), transfer learning	Radiographs, whole-body CT	Multicenter series	Mostly internal	Accuracy 81–96%; sensitivity up to 98% per study; radiologist-level performance	Grad-CAM / saliency maps in some

Table 2: Study-level characteristics of principal healing- and outcome-prediction studies

Study (ref)	Design / setting	Input data / modality	Algorithm(s)	Prediction target	Key performance	Validation
Razavi et al. 2025 [18]	Simulation study (validated rodent FE model)	Implant configuration, loading regime, biological parameters	SVR, RF, XGBoost, MLP, CNN, RNN, LSTM	Callus stiffness / healing trajectory	High agreement with FE ground truth; robust at 50% training data	Internal (unseen combinations); SHAP interpretation
Liu et al. 2023 [19]	Simulation study (3-D DRF healing model)	Fracture geometry, gap size, loading, healing time (3,600 cases)	Multiple ML incl. cubic SVM, trilayered ANN	Healing outcome by post-injury stage	Cubic SVM best early; ANN best late	Internal
North et al. 2024 [13]	Retrospective cohort, single center (USA)	Step-level insole ground-reaction-force features	Decision tree + LASSO feature selection → logistic regression	Days-to-healing (30-day window)	≈76% accuracy / precision / recall / F1	10-fold and leave-one-out CV
Yu et al. 2025 [12]	Animal imaging study (rat micro-CT)	2,448 micro-CT images + bone morphometric parameters	VM-TE-UNet segmentation + early-diagnosis model	Early non-union screening (day 14)	Dice 0.809; 95th Hausdorff 13.1	Internal
Tümen et al. 2025 [24]	Retrospective, 346 scaphoid non-union patients	Pre- and post-operative radiographs (imaging only)	TensorFlow-based deep learning	Union after revision surgery	≈93.6% accuracy	Internal
Suter et al. 2024 [8]	Cohort study (humeral	Serial radiographs;	Radiographic scoring +	Humeral shaft non-	RUSHU ≤8: RR 4.3 (95%	Internal

	shaft fractures)	RUSHU at 6 weeks	statistical analysis	union	CI 2.7–6.9); cut-off ≤7: 82% specificity / 63% sensitivity	
Mosfeldt et al. 2024 [27]	Registry cohort (admission parameters)	Demographics, biochemistry, comorbidities	Multiple ML classifiers	1-, 3-, 6-, 12-month mortality	Good-to-excellent discrimination	Internal
Kitcharanant et al. 2022 [28]	Registry-derived cohort (Thailand)	Demographics and comorbidities (CCI, heart disease, BMI, dementia, lung disease)	Seven ML classifiers (GB, RF, ANN, LR, NB, SVM, KNN)	1-year mortality after fragility hip fracture	Gradient boosting and random forest best; web application deployed	Internal
Li et al. 2021 [29]	Retrospective cohort	Preoperative clinical variables	Novel machine-learning algorithm	Mortality risk after hip fracture surgery	Improved risk stratification vs conventional approaches	Internal
Shuchami et al. 2026 [30]	Multicenter registry (Israel)	Electronic health record variables	Manual vs automated ML/DL pipelines	1-year mortality in elderly hip fracture	Automated pipelines competitive with manual modeling	Internal
Dijkstra et al. 2025 [14]	National audit (Netherlands), 74,396 cases	Preoperative demographic and clinical variables	Logistic regression and ML algorithms	Short- and long-term mortality for shared decision-making	Adequate performance for decision support	Internal (split-sample)
Wang 2025 [31]	Retrospective administrative cohort	Sociodemographic and comorbidity data	Naïve Bayes vs ensemble ML	In-hospital mortality after osteoporotic hip fracture	Naïve Bayes on par with other ML; CKD, cardiovascular and bone-metabolism comorbidities most influential	Internal; interpretable by design

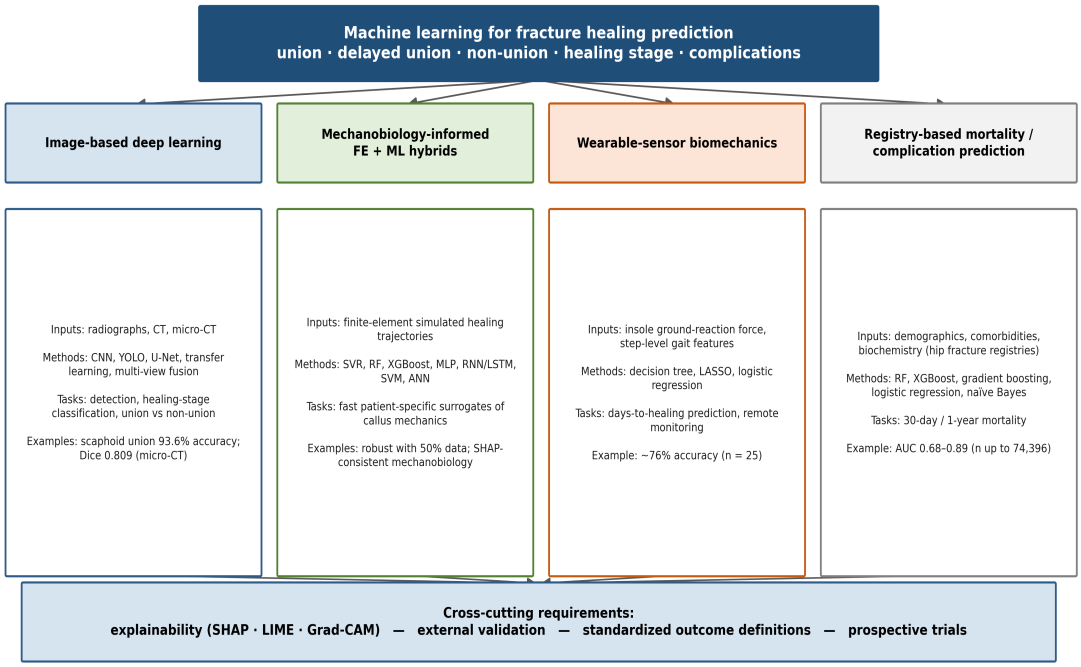

Figure 2: Taxonomy of machine learning approaches to fracture healing prediction, with representative inputs, methods, tasks, and results.

Model Interpretability and Explainable Artificial Intelligence: A recurring barrier to clinical adoption of ML-based prediction models, across medicine broadly and orthopedic trauma specifically, is the "black-box" nature of complex algorithms such as deep neural networks and gradient-boosted ensembles [15,35]. Explainable artificial intelligence (XAI) techniques — most prominently SHapley Additive exPlanations (SHAP), which quantify each input feature's contribution to an individual prediction using a game-theoretic framework — have become the dominant approach for addressing this gap in clinical prediction modeling [35]. Within the fracture healing literature specifically, SHAP analysis of the mechanobiology-informed ML surrogate model discussed above revealed biologically coherent relationships, such as the positive influence of screw number on central callus stiffness and the negative influence of excessive mechanical loading on outer callus formation, thereby lending clinical and mechanistic credibility to model outputs beyond raw predictive accuracy [18]. Similarly, interpretable modeling of osteoporotic hip-fracture in-hospital mortality identified chronic kidney disease, cardiovascular comorbidities, and markers of bone metabolism as the most influential predictors, aligning with established clinical intuition and supporting clinician trust in model outputs [31].

More broadly, comprehensive reviews of explainable AI in healthcare identify SHAP as the most frequently applied interpretability technique for structured clinical data, alongside Local Interpretable Model-agnostic Explanations (LIME) for tabular data and gradient-weighted class activation mapping (Grad-CAM) for imaging-based models [35]. The integration of such interpretability methods is likely to be a prerequisite, rather than an optional enhancement, for the clinical adoption of fracture healing prediction models, given the safety-critical nature of decisions — such as revision surgery timing — that such predictions may inform.

Challenges and Limitations: Despite encouraging performance metrics across the reviewed literature, several methodological limitations recur throughout this body of work and constrain its current clinical readiness. First, most studies are retrospective and derived from single-center cohorts with sample sizes ranging from several dozen to a few thousand patients, raising concerns regarding overfitting and limited generalizability to different patient populations, imaging equipment, and clinical protocols [13,24,26]. Second, external validation — testing a model developed in one population on an independent, geographically or temporally distinct cohort — remains infrequent across the reviewed studies, and where reported, model performance often degrades relative to internal validation results [10,26]; in a comprehensive review of AI in orthopedic trauma, only 14.5% of studies underwent external validation and just 3.2% reported prospective clinical validation [11]. Third, outcome definitions for delayed union and non-union vary considerably between studies, with inconsistent time thresholds and radiographic or clinical criteria, which complicates direct comparison of reported model performance and precludes meaningful meta-analytic pooling [26].

Fourth, class imbalance is a persistent issue, since non-union represents a minority outcome in most fracture cohorts, which can bias model training toward the majority "union" class unless specifically addressed through resampling or cost-sensitive learning strategies.

Fifth, several imaging-based studies rely on animal models (notably rodent micro-CT) whose healing timelines and biology differ substantially from human fracture repair, necessitating caution when extrapolating findings to clinical practice [12].

Sixth, model calibration and clinical net benefit (e.g., decision-curve analysis) are infrequently reported, so well-discriminating models may still be poorly suited to threshold-based clinical decisions [14,26].

Seventh, regulatory and governance considerations — including software-as-a-medical-device classification, liability, and prospective safety monitoring — remain largely unaddressed, as do fairness and bias concerns arising from datasets skewed by age, sex, fracture etiology, implant type, or imaging vendor.

Finally, integration into real-world clinical workflows remains largely unaddressed; few studies describe prospective deployment, clinician-in-the-loop evaluation, or the practical infrastructure required to incorporate model outputs into routine orthopedic decision-making.

Future Directions: Several research priorities emerge from this synthesis (Figure 3). Multi-center prospective data collection and sharing initiatives are needed to build sufficiently large, diverse, and well-annotated datasets capable of supporting robust external validation, mirroring efforts already underway in hip-fracture mortality prediction through national audit datasets [14]. Standardized outcome definitions and reporting frameworks — such as adaptation of the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) and CLAIM (Checklist for Artificial Intelligence in Medical Imaging) guidelines specifically for fracture healing prediction — would substantially improve comparability across studies [26], and formal risk-of-bias appraisal using PROBAST should accompany future systematic evaluations. Multi-modal data fusion, combining radiographic imaging, wearable-sensor biomechanical data, and structured clinical variables within unified models, represents a promising direction given the complementary information captured by each modality [13,18]. Continued development of mechanobiology-informed ML surrogates offers the prospect of real-time, patient-specific simulation of healing trajectories at the point of care, without the computational burden of full finite-element analysis [18,19]. Finally, prospective, ideally multi-center, clinical validation studies — including trials assessing whether model-guided intervention improves patient outcomes relative to standard care — will be essential to move this field from retrospective proof-of-concept toward routine clinical utility.

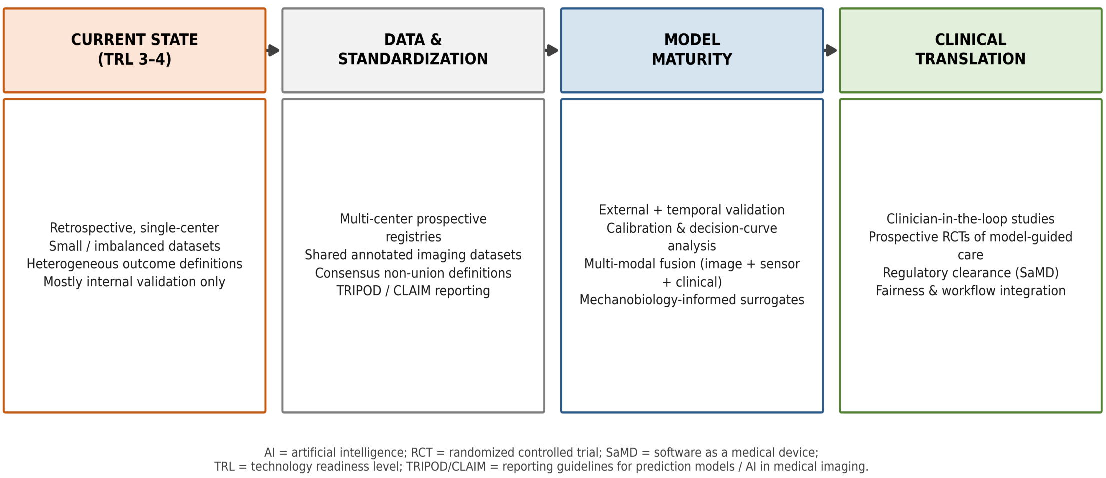

Figure 3: From current barriers to clinical translation: a research roadmap for ML-based fracture healing prediction.

Conclusion

Machine learning and deep learning methods are increasingly being applied to the prediction of fracture healing outcomes, spanning radiographic and micro-CT image analysis, mechanobiology-informed simulation surrogates, wearable-sensor biomechanical monitoring, and clinical registry-based mortality and complication prediction. Reported performance across these domains is frequently strong, and the growing incorporation of explainable AI techniques such as SHAP is helping to bridge the interpretability gap that has historically limited clinician trust in algorithmic outputs.

Nonetheless, the field remains predominantly at a retrospective, single-center, proof-of-concept stage — approximately technology readiness level 3–4 — with limited external validation and inconsistent outcome definitions constraining current clinical applicability. Addressing these gaps through multi-center data sharing, standardized reporting, multi-modal model development, and prospective clinical validation will be essential to realize the potential of machine learning as a tool for earlier, more objective, and more individualized prediction of fracture healing in routine orthopedic trauma care.

Until such validation is achieved, no currently published model should be used for standalone clinical decision-making; ML tools should be regarded as research adjuncts that complement, not replace, clinical and radiographic assessment.

Declarations

Author contributions (CRediT): Narendra B S — Conceptualization, Writing – original draft. Prajwal J — Literature search, Data curation, Writing – original draft. Pramod G — Supervision, Writing – review & editing. All authors read and approved the final manuscript.

Use of generative AI: Artificial intelligence–assisted tools were used only for language refinement during manuscript preparation; the authors take full responsibility for the accuracy and integrity of all content.

Ethics approval and consent to participate: Not applicable — this is a narrative review of previously published literature.

Data availability: Not applicable — no new data were generated or analyzed in this study.

Abbreviations: AI, artificial intelligence; ANN, artificial neural network; AUC, area under the receiver operating characteristic curve; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CNN, convolutional neural network; CT, computed tomography; FE, finite element; GRF, ground reaction force; LASSO, least absolute shrinkage and selection operator; LIME, Local Interpretable Model-agnostic Explanations; LSTM, long short-term memory; micro-CT, micro-computed tomography; ML, machine learning; MLP, multilayer perceptron; PROBAST, Prediction model Risk Of Bias Assessment Tool; RF, random forest; RNN, recurrent neural network; RUSH, Radiographic Union Score for Hip fractures; RUSHU, Radiographic Union Score for HUmeral fractures; RUST, Radiographic Union Score for Tibial fractures; SANRA, Scale for the Assessment of Narrative Review Articles; SHAP, SHapley Additive exPlanations; SVM, support vector machine; SVR, support vector regression; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; XAI, explainable artificial intelligence.

References

1. Zura R, Xiong Z, Einhorn T, Watson JT, Ostrum RF, Prayson MJ, et al. Epidemiology of fracture nonunion in 18 human bones. JAMA Surg. 2016;151(11):e162775. doi:10.1001/jamasurg.2016.2775.
2. Molitoris KH, Huang M, Baht GS. Osteoimmunology of fracture healing. Curr Osteoporos Rep. 2024;22(3):330–339. doi:10.1007/s11914-024-00869-z.
3. Whelan DB, Bhandari M, Stephen D, Kreder H, McKee MD, Zdero R. Development of the radiographic union score for tibial fractures for the assessment of tibial fracture healing after intramedullary fixation. J Trauma. 2010;68(3):629–632.
4. Kooistra BW, Dijkman BG, Busse JW, Sprague S, Schemitsch EH, Bhandari M. The radiographic union scale in tibial fractures: reliability and validity. J Orthop Trauma. 2010;24(Suppl 1):S81–S86.
5. Chiavaras MM, Bains S, Choudur H, Parasu N, Jacobson J, Ayeni O, et al. The Radiographic Union Score for Hip (RUSH): the use of a checklist to evaluate hip fracture healing improves agreement between radiologists and orthopedic surgeons. Skeletal Radiol. 2013;42(8):1079–1088. doi:10.1007/s00256-013-1605-8.
6. Frank T, Osterhoff G, Sprague S, Garibaldi A, Bhandari M, Slobogean GP; FAITH Investigators. The Radiographic Union Score for Hip (RUSH) identifies radiographic nonunion of femoral neck fractures. Clin Orthop Relat Res. 2016;474(6):1396–1404. doi: 10.1007/s11999-015-4680-4.
7. Oliver WM, Smith TJ, Nicholson JA, Molyneux SG, White TO, Clement ND, et al. The Radiographic Union Score for HUmeral fractures (RUSHU) predicts humeral shaft nonunion. Bone Joint J. 2019;101-B(10):1300–

1306. doi:10.1302/0301-620X.101B10.BJJ-2019-0304.R1.
8. Suter C, Mattila H, Ibounig T, Sumrein BO, Launonen A, Järvinen TLN, et al. Prediction of humeral shaft fracture healing using the Radiographic Union Score for HUmeral Fractures (RUSHU). Bone Jt Open. 2024;5 (11):962–970. doi:10.1302/2633-1462.511.BJO-2024-0134.R1.
9. Wurm LM, Ertel W, Laue D. Artificial intelligence in orthopedic trauma surgery: a scoping review of current applications and research gaps. Arch Orthop Trauma Surg. 2026;146:122. doi:10.1007/s00402-026-06276-6.
10. Dijkstra H, van de Kuit A, de Groot T, Canta O, Groot OQ, Oosterhoff JHF, Doornberg JN; Machine Learning Consortium. Systematic review of machine-learning models in orthopaedic trauma: an overview and quality assessment of 45 studies. Bone Jt Open. 2024;5(1):9–19.
11. Misir A. Artificial intelligence in orthopedic trauma: a comprehensive review. Injury. 2025; 56(8):112570. doi:10.1016/j.injury.2025.112570.
12. Yu H, Mu Q, Wang Z, Guo Y, Zhao J, Wang G, et al. A study on early diagnosis for fracture non-union prediction using deep learning and bone morphometric parameters. Front Med. 2025;12:1547588. doi:10.3389/fmed.2025.1547588.
13. North K, Simpson G, Geiger W, Cizik A, Rothberg D, Hitchcock R. Predicting the healing of lower extremity fractures using wearable ground reaction force sensors and machine learning. Sensors. 2024;24(16):5321. doi:10.3390/s24165321.
14. Dijkstra H, Parsons CS, van Bremen HE, Willems HC, de Hond AAH, van Munster BC, Doornberg JN, Oosterhoff JHF. Machine learning-based prediction of short- and long-term mortality for shared decision-making in older hip fracture patients: the Dutch Hip Fracture Audit algorithms in 74,396 cases. Acta Orthop. 2025;96:521–528. doi:10.2340/17453674.2025.44248.
15. Hendrix N, Scholten E, Vernhout B, Bruijnen S, Maresch B, de Jong M, et al. Development and validation of a convolutional neural network for automated detection of scaphoid fractures on conventional radiographs. Radiol Artif Intell. 2021;3(4):e200260. doi:10.1148/ryai.2021200260.
16. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. 2019;29(10):5469–5477. doi:10.1007/s00330-019-06167-y.
17. Ghiasi MS, Chen J, Vaziri A, Rodriguez EK, Nazarian A. Bone fracture healing in mechanobiological modeling: a review of principles and methods. Bone Reports. 2017; 6:87–100. doi:10.1016/j.bonr.2017.03.002.
18. Razavi AH, Hemmati M, Nafisi N, Mirahmadi A, Laiwala S, Keko M, Vaziri A, Rodriguez EK, Nazarian A. Mechanobiology-guided machine learning models for predicting long bone fracture healing across diverse scenarios. Comput Biol Med. 2025;196(Pt A):110683. doi:10.1016/j.compbiomed.2025.110683.
19. Liu X, Miramini S, Patel M, Ebeling P, Liao J, Zhang L. Development of numerical model-based machine learning algorithms for different healing stages of distal radius fracture healing. Comput Methods Programs Biomed. 2023;233:107464. doi:10.1016/j.cmpb.2023.107464.
20. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell. 2019;1(1):e180001. doi:10.1148/ryai.2019180001.
21. Lu J, Ding J, Wang P, Mi B, Zhou P, Zheng F. A deep learning framework for automated fracture detection and localization in radiographic images using convolutional neural networks. Front Med. 2026;13:1806133. doi:10.3389/fmed.2026.1806133.
22. Jha G, Dhanjal R, Malasani S, Karunakaran AV. Deep learning in scaphoid fracture detection and healing prediction: a systematic review of artificial intelligence applications in hand surgery. Cureus. 2025;17(11):e97900. doi:10.7759/cureus.97900.
23. Zdolsek G, Chen Y, Bögl HP, Wang C, Woisetschläger M, Schilcher J. Deep neural networks with promising diagnostic accuracy for the classification of atypical femoral fractures. Acta Orthop. 2021;92(4):394–400. doi:10.1080/17453674.2021.1891512.
24. Tümen L, Medved F, Rachunek-Medved K, Han Y, Saul D. Deep learning in scaphoid nonunion treatment. J Clin Med. 2025;14(6):1850. doi:10.3390/jcm14061850.
25. Yüce A, Yaşar H, Misir A. Can fracture non-union be predicted using deep learning? Explor Musculoskelet Dis. 2025;3:100790. Available from:https://www.explorationpub.com/Journals/emd/Article/100790.
26. Yin F, Wang X, Luo G, Deng Z, Chen W, Zhang Y. Performance of prediction models for delayed union and nonunion after fracture: a systematic review and meta-analysis. BMC Musculoskelet Disord. 2026;27:487. doi:10.1186/s12891-026-09941-4.
27. Mosfeldt M, Jørgensen HL, Lauritzen JB, Jansson KÅ. Development and internal

validation of a multivariable prediction model for mortality after hip fracture with machine learning techniques. Calcif Tissue Int. 2024;114(6):568–582. doi:10.1007/s00223-024-01208-1.
28. Kitcharanant N, Chotiyarnwong P, Tanphiriyakun T, Vanitcharoenkul E, Mahaisavariya C, Boonyaprapa W, Unnanuntana A. Development and internal validation of a machine-learning-developed model for predicting 1-year mortality after fragility hip fracture. BMC Geriatr. 2022;22:451. doi:10.1186/s12877-022-03152-x.
29. Li Y, Chen M, Lv H, Yin P, Zhang L, Tang P. A novel machine-learning algorithm for predicting mortality risk after hip fracture surgery. Injury. 2021;52(6):1487–1493. doi:10.1016/j.injury.2020.12.008.
30. Shuchami A, Glebov M, Katsin M, Portnoy Y, Berkenstadt H, Orkin D, Lazebnik T. Comparing manual vs. automated machine learning and deep learning models for predicting one-year mortality in elderly hip fracture patients. Front Med. 2026;13:1804645. doi:10.3389/fmed.2026.1804645.
31. Wang JWD. Naïve Bayes is an interpretable and predictive machine learning algorithm in predicting osteoporotic hip fracture in-hospital mortality compared to other machine learning algorithms. PLOS Digit Health. 2025;4(1):e0000529. doi:10.1371/journal.pdig.0000529.
32. Adebayo O, Bhuiyan ZA, Ahmed Z. Exploring the effectiveness of artificial intelligence, machine learning and deep learning in trauma triage: a systematic review and meta-analysis. Digit Health. 2023;9:20552076231205736. doi: 10.1177/20552076231205736.
33. Vaishya R, Saiyed A, Vaish A, Botchu R, Iyengar KP, et al. Artificial intelligence and decision support in orthopaedic emergencies: a review. Indian J Orthop. 2026. doi:10.1007/s43465-026-01760-y.
34. Geda MW, Tang YM, Lee CKM. Applications of artificial intelligence in orthopaedic surgery: a systematic review and meta-analysis. Eng Appl Artif Intell. 2024; 133:108326. doi:10.1016/j.engappai.2024.108326.
35. Zhang K, Wang D, Lin F, Xie J, Zhou W. A comprehensive review of explainable artificial intelligence in healthcare: methods, evaluation, and clinical integration. iScience. 2026. doi:10.1016/j.isci.2026.115026.