Journal of Orthopaedic Surgery and Research

RESEARCH

Open Access

Machine learning-based prediction of the necessity for the surgical treatment of distal radius fractures

Jongmin Lim[1†], Sehun Chang[1†], Kwangsu Kim[1], Hee Jin Park[2], Eugene Kim[3] and Seok Woo Hong[3*]

Abstract

Background Treatments for distal radius fractures (DRFs) are determined by various factors. Therefore, quantitative or qualitative tools have been introduced to assist in deciding the treatment approach. This study aimed to develop a machine learning (ML) model that determines the need for surgical treatment in patients with DRFs using a ML model that incorporates various clinical data concatenated with plain radiographs in the anteroposterior and lateral views.

Methods Radiographic and clinical data from 1,139 patients were collected and used to train the ML models. To analyze and integrate data effectively, the proposed ML model was mainly composed of a U-Net-based image feature extractor for radiographs, a multilayer perceptron based clinical feature extractor for clinical data, and a final classifier that combined the extracted features to predict the necessity of surgical treatment. To promote interpretability and support clinical adoption, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to provide visual insights into the radiographic data. SHapley Additive exPlanations (SHAP) were utilized to elucidate the contributions of each clinical feature to the predictions of the model.

Results The model integrating image and clinical data achieved accuracy, sensitivity, and specificity of 92.98%, 93.28%, and 92.55%, respectively, in predicting the need for surgical treatment in patients with DRFs. These findings demonstrate the enhanced performance of the integrated model compared to the image-only model. In the Grad-CAM heatmaps, key regions such as the radiocarpal joint, volar, and dorsal cortex of the radial metaphysis were highlighted, indicating critical areas for model training. The SHAP results indicated that being female and having subsequent or concomitant fractures were strongly associated with the need for surgical treatment.

Conclusions The proposed ML models may assist in assessing the need for surgical treatment in patients with DRFs. By improving the accuracy of treatment decisions, this model may enhance the success rate of fracture treatments, guiding clinical decisions and improving efficiency in clinical settings.

Keywords Fracture management, Surgical decision-making, Clinical data integration, U-Net, Grad-CAM, SHAP

[†]Jongmin Lim and Sehun Chang contributed equally to this work and share co-first authorship.

*Correspondence:
Seok Woo Hong
poisoxic@naver.com

[1]Department of Computer Science and Engineering, Sungkyunkwan University College of Computing and Informatics, Suwon, South Korea
[2]Department of Radiology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, South Korea
[3]Department of Orthopedic Surgery, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, South Korea

Introduction

Distal radius fractures (DRFs) are the most common types of upper extremity fractures [1], with their prevalence increasing as society enters a super-aging era [2]. They commonly occurs in older patients with osteoporosis or in postmenopausal women [3]. Treatment options for DRFs can generally be divided into conservative, such as cast immobilization, and surgical methods, including open reduction and internal fixation. Deciding on the treatment method for patients with DRFs is highly dependent on whether the degree of fracture displacement and joint incongruency observed on plain radiographs and computed tomography scans falls within an acceptable range [1]. However, the decision to pursue surgical treatment is often influenced by factors such as the clinical experience of the orthopedic surgeon, surgical demands of the patients, and other factors such as age, gender, and pre-existing health conditions [4].

Older patients with DRFs often have various underlying conditions [5], which can lead to debates regarding the appropriate treatment approach. Orthopedic surgeons frequently encounter difficulties in deciding whether to offer surgical treatment to correct bone misalignment and restore joint congruency, despite the risks of complications associated with surgery such as anesthesia side effects, reduced joint mobility, and postoperative infections, or to opt for conservative treatment to avoid the aforementioned risks while accepting the risk of wrist deformity or joint incongruency. Furthermore, determining which treatment option is the best from patients' perspective, particularly in clinical settings, is often difficult. This is a common issue in not only DRFs but also fractures in other areas.

In recent years, many studies have reported findings using machine learning (ML) models in orthopedics [6]. These studies have included fracture detection and classification, bone age calculation, assessment of the arthritis severity, determination of implant types, and evaluation of fracture risks [7]. These advancements may have contributed to improving diagnostic accuracy and treatment planning in orthopedics. In particular, previous ML research on fracture detection and classification has provided a strong basis for identifying and categorizing various fracture types, which is crucial for developing appropriate treatment plans [8, 9, 10, 11]. Furthermore, ML studies on osteoarthritis or rheumatoid arthritis have emphasized the influence of patient demographics, such as age and underlying health conditions, on treatment decisions [12]. In fracture research, many studies have focused on developing ML models to detect fractures and classify fracture types in areas such as the spine, hip, humerus, and scaphoid [13, 14, 15, 16, 17, 18]. Moreover, recent studies on the development of ML models for detecting DRFs and predicting the risk factors for postoperative complications following open reduction and internal fixation surgery for DRFs have been published [19, 20, 21]. However, no ML models have been developed for determining treatment plans after diagnosing fractures, including DRFs.

Recently, numerous studies in the field of medical ML have emphasized the training of models through the integration of multimodal data, rather than relying on a single data type [22]. Integrating different data forms may contribute to the development of more robust and accurate models, significantly enhancing their potential for application in diagnosis and treatment [22]. Several studies have investigated the use of combined image and clinical data for training ML models, among the different approaches to data integration [23, 24, 25]. However, the integration of image data and clinical data has not been applied to ML research for fracture management in orthopedics.

Therefore, the objective of this study was to develop a ML model that determines the need for surgical treatment in patients with DRFs by incorporating clinical data concatenated with radiographs in the anteroposterior (AP) and lateral views. Through this, we aimed to assist in determining the treatment option for DRFs.

Materials and methods

Datasets

The study protocol was approved by the institutional review board of a university hospital (IRB No. KBSMC 2023-01-044) and the requirement for obtaining informed consent was waived. A total of 1,766 patients aged ≥ 18 years who received conservative treatment for DRFs between January 2001 and December 2023 were initially included in this study. The inclusion criteria were as follows: (1) plain radiographs taken in AP and lateral views with a resolution of ≥ 300 dpi, (2) plain radiographs taken before fracture reduction, (3) closed growth plate, and (4) accessible clinical information in the electronic medical records. Collected clinical data included the patient's age, gender, injured side, body mass index (BMI), and the presence of concomitant or subsequent fractures. Patients whose plain radiographs could not be analyzed because of bone obstruction by a cast or splint were excluded. Ultimately, radiographs and clinical data from 1,139 patients were analyzed in the study. The plain radiographs were obtained using GC85A (Samsung Electronics Co., Ltd., Suwon, Republic of Korea) plain radiography equipment, and the images were extracted in JPEG format. The original images had a pixel resolution of 1452 × 1816 and a size of 0.085 mm/pixel.

Image labeling

Three upper extremity musculoskeletal system experts (SWH, EK, and HJP) individually assessed the need for

Table 1 Data augmentation strategy and details

Augmentation Type	Function	Parameter	Augmentation Details	Probability
Cropping	Random Resized Crop	Crop Area Scale	A value is randomly selected between -50% and 50% of the original image area is cropped and resized into output size (output size = 512 × 512 × 3)	1.0
Flip	Horizontal Flip	Flip Direction	The image is flipped horizontally (mirrored along the vertical axis)	0.5
Color Adjustment	Color Jitter	Brightness	A value is randomly selected between -50% and 50% of the original value	0.5
		Contrast	A value is randomly selected between -50% and 50% of the original value	0.5
		Saturation	A value is randomly selected between -50% and 50% of the original value	0.5
		Hue	A value is randomly selected between -10% and 10% of the original value	0.5
	Solarization	Threshold	Any pixel with a value above 128 of the images (on a scale of 0–255) are inverted (original value range = 0–255)	0.5
Brightness/Contrast	Random Brightness Contrast	Brightness Limit	A value is randomly selected between -20% and 20% of the original value	0.2
Optical Distortion	Optical Distortion	Distort Limit	A value is randomly selected between -5% and 5% of the original value	0.5
		Shift Limit	A value is randomly selected between -5% and 5% of the original value	

surgical treatment using plain radiographs and the clinical information of each patient, with each expert blinded to the others. No formal expert training or calibration session was conducted before image labeling. Each case was assessed twice by three experts. If all six assessments matched, the score was accepted; consensus discussion was employed to resolve disagreements. The need for surgical treatment by each patient was evaluated as either "surgery required (yes)" or "surgery not required (no)," and this information was annotated to each patient's case for supervised ML. The reliability of the assessments regarding the need for surgical treatment was evaluated using the intra-class correlation coefficient (ICC). First, inter-observer reliability was evaluated using ICC with two-way random effects and absolute agreement with the mean of the multiple measurements model (ICC [2, k]). Then, intra-observer reliabilities were evaluated by an ICC using two-way random effects and absolute agreement with the single measurement model (ICC [2, 1]). Intra-observer reliability was determined by evaluating each quality score at 2-week intervals to ensure independent assessments.

Data composition and augmentation

The dataset was obtained through random sampling and divided into training, validation, and test sets. The sets were distributed in an approximate ratio of 16:4:5. Therefore, 728, 183, and 228 cases were assigned to the training, test, and validation sets, respectively, and 1456, 366, and 456 images were assigned to the training, test, and validation sets, respectively. All radiographic images were resized to 512 × 512 pixels and normalized. To enhance model generalization, the images were subjected to random augmentation. Each augmentation technique was applied independently and probabilistically based on its specified probability, allowing multiple transformations to be applied simultaneously to a single image. Table 1 shows a detailed summary of the parameter settings and application probabilities for each augmentation. To avoid bias in performance evaluation, these augmentations were applied only during training and not during validation or testing [26]. Additionally, to ensure numerical stability, the clinical data features were normalized by min–max scaling [27].

Table 2 Hardware and software environment specification

Hardware	Product and manufacturer name	Detailed features
CPU	Intel Xeon Gold 6442Y	24-core, 4.0 GHz boost clock
RAM	SAMSUNG DDR5 PC5-4800	128 GB
GPU	NVIDIA RTX4090	GDDR6X, 24 GB
Software	**Detailed features**	
Operating system	Ubuntu 22.04 with CUDA 11.5/CuDNN 8.1.0	
Machine learning framework	PyTorch version 2.5.1	

CPU, Central processing unit; RAM, Random access memory; GPU, Graphic processing unit

ML architecture and details

ML models were developed using Pytorch (version 2.5.1, Meta AI, CA, USA) ML framework and the Python (version 3.10.12) programming language. The specific developmental environments are summarized in Table 2. Model training was performed by utilizing only radiographic images (image-only model) and incorporating clinical data to enhance predictive performance (model integrating image and clinical data).

Figure 1 shows the U-Net based image feature extractor for the radiographic images [28]. It processes the AP and lateral radiographic views of each patient to capture pixel-level hierarchical features from the plain radiographs. As shown in Fig. 1, these extractors follow a fully

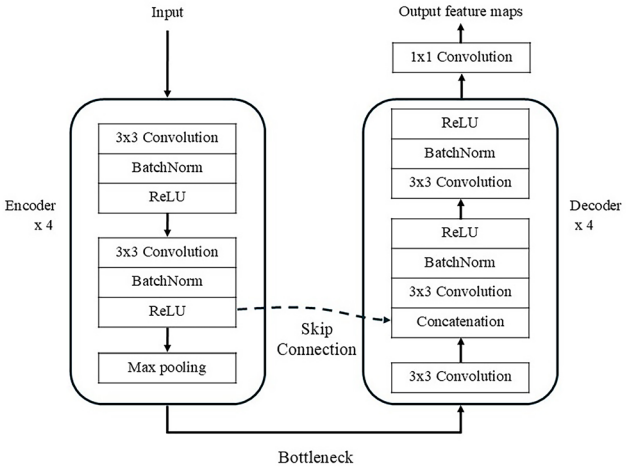

Fig. 1 U-net based image feature extractor for plain radiograph images

convolutional encoder–decoder architecture with skip connections and a bottleneck layer. The encoder progressively extracts multiscale structural features through convolutional layers that include batch normalization and rectified linear unit (ReLU) activation, followed by max pooling for spatial downsampling. The bottleneck layer refines these representations by extracting high-level semantic features, thereby enhancing the ability of the model to differentiate structural patterns in radiographic images. The decoder restores spatial details through skip connections, which create direct pathways between the corresponding encoder layers to recover fine-grained anatomical structures that may have been lost during encoding. Ultimately, a 1×1 convolutional layer compresses the high-dimensional feature space into a compact representation while preserving spatial integrity, ensuring its suitability for classification (Supplementary Fig. 1). The extracted feature maps from the AP and lateral radiographic views were then concatenated and passed into the final classifier, enabling the model to analyze fractures from multiple perspectives (Fig. 2A, image-only model).

The image and clinical data-integrating model was trained through the following steps. First, as in the image-only model, the radiographics were processed through a U-Net-based image feature extractor. Next, the clinical data were analyzed using a clinical feature extractor, implemented as a multilayer perceptron consisting of two fully connected layers and ReLU activation [29]. The features extracted from both modalities were concatenated and passed into the final classifier, allowing the model to leverage complementary information to improved predictive performance (Fig. 2B).

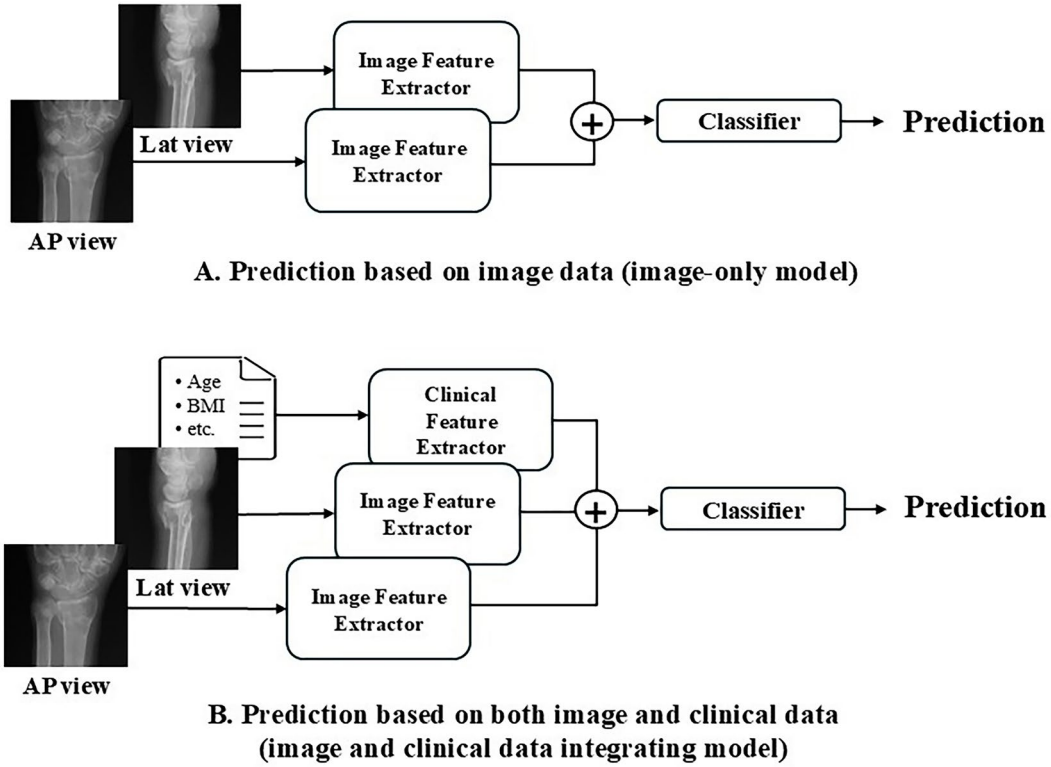

Fig. 2 Overview of the prediction models. **A.** Model utilizing only radiographic image data as input. Image features are extracted through an image feature extractor and fed into a classifier to generate predictions. **B.** Model integrating both radiographic images and clinical data (e.g., age and BMI). Clinical features were extracted using a clinical feature extractor and concatenated with image features before being passed to the classifier. Integrating image and clinical data leverages complementary information, potentially improving predictive performance

Table 3 Performance indicator formulas

Performance indicator	Formulas
Accuracy	True positive + True negative / True positive + False positive + True negative + False negative
Sensitivity	True positive / True positive + False negative
Specificity	True negative / True negative + False positive
Positive predictive value	True positive / True positive + False positive
Negative predictive value	True negative / True negative + False negative

The proposed model was trained using cross-entropy loss and optimized with the Adam optimizer, where the learning rate, first moment decay rate, second moment decay rate, and weight decay were set to 1.0×10^{-5}, 0.9, 0.999, and 1.0×10^{-6}, respectively. A batch size of 8 was used, and training proceeded for 200 epochs, with early stopping based on validation loss with a patience parameter of 10 to prevent overfitting. To identify the regions in the images that contributed most significantly to the model's prediction, the Gradient-weighted Class Activation Mapping (Grad-CAM) method was employed. This method provides a visual interpretation of model's prediction by highlighting the most relevant regions in the image [30]. A masking-based training strategy was implemented to assess the actual contribution of the regions highlighted by Grad-CAM to the decision-making process of the model. We conducted additional model training by excluding the regions highlighted by the Grad-CAM as follows: (1) A machine learning model was initially trained using the original training set images. (2) The trained model was used to generate Grad-CAM heatmaps for the training set images. (3) Each Grad-CAM heatmap was normalized to a range of [0, 1]. (4) Regions with high activation values exceeding a given threshold were masked. Pixels with values above the designated threshold (e.g., 0.5, 0.8, and 1.0) were set to zero, whereas those below the threshold retained their original values. (5) New models were subsequently trained using the modified training set images corresponding to each threshold. The performance of the retrained models was evaluated using the unmodified test set to determine the contribution of the regions highlighted by Grad-CAM to the decision-making process of the models. Additionally, the SHapley Additive exPlanations (SHAP) were utilized to quantify the relative contributions of each clinical feature to the model predictions [31]. It provides feature-specific explanations, enabling a clearer understanding and more informed clinical decision-making based on patient-specific data.

Statistical analysis

The diagnostic performances of the ML models were evaluated by calculating the area under the curve (AUC) from the receiver operating characteristic (ROC) curve, along with metrics such as accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. The calculated performance indicators are presented in Table 3.

Table 4 Clinical characteristics of the studied patients

Characteristics	Surgery not required	Surgery required	Total	*P* value
Participants	480	659	1139	
Mean age (yr)	54.88 ± 16.51 (19–93)	63.09 ± 15.93 (19–96)	59.63 ± 16.67	< 0.001*
Gender (Men / Women)	190 (39.6%) / 290 (60.4%)	158 (24.0%) / 501 (76.0%)	348 / 791	< 0.001*
Affected sides (Right / Left)	329 (50.0%) / 329 (50.0%)	229 (34.7%) / 250 (65.3%)	558 / 579	0.465
Body mass index (kg/m^2)	24.62 ± 3.74 (16.9–46.6)	23.93 ± 3.52 (14.9–35.7)	24.22 ± 3.63	0.002*
Presence of subsequent or concomitant fracture (No / Yes)	374 (77.9%) / 106 (22.1%)	532 (80.7%) / 127 (19.3%)	906 / 233	0.245

Descriptive values are shown as mean ± standard deviation (range) or number of cases (proportion (%)); Statistical differences between the two groups were analyzed using the independent t-test for continuous variables and the Chi-square test for categorical variables

*Adjusted $P < 0.01$ by chi-square test and independent t-test

The McNemar test was used to evaluate the statistical significance of the difference in classification performance between the image-only model and the image and clinical data integrating model; both have a dichotomous dependent variable. To assess whether a statistically significant difference exists between the AUCs of the two models, DeLong's test was employed. The Shapiro–Wilk and Kolmogorov–Smirnov normality tests revealed that the clinical data were normally distributed. Therefore, the chi-square or independent t-test was used to assess whether a significant difference exists in the clinical data between patients who required surgery and those who did not. Because five simultaneous comparisons were performed in the chi-square and independent t-tests, a Bonferroni correction was applied. Statistical significance was defined as $P < 0.05$ for the McNemar test and DeLong's test, and $P < 0.01$ (0.05/5) for the chi-square and independent t-tests. All statistical analyses were performed using scikit-learn Python library (version 1.5.0).

Results

Table 4 presents the clinical characteristics of the study participants, and Table 5 shows the intra-observer [ICC (2, 1)] and inter-observer [ICC (2, k)] reliabilities for the expert assessments. With ICC values ≥ 0.9 indicating excellent reliability [32], the intra- and inter-observer reliabilities for the quality assessments were acceptable.

Figure 3A and B illustrate the confusion matrices for the image-only model and the image and clinical data integrating model, respectively. The image-only model

Table 5 Intra-observer and inter-observer reliabilities of expert assessments

Experts	ICC (2, 1) for intra-observer reliability	ICC (2, k) for inter-observer reliability
SWH	0.962	0.965
HJP	0.971	
EK	0.960	

ICC (2, 1), intra-class correlation coefficient (ICC) using 2-way random effects and absolute agreement with the single measurement model; ICC(2, k), ICC using 2-way random effects and absolute agreement with the mean of multiple measurements model

Table 6 Diagnostic performances of current machine-learning models

Model type Performance indicator	Image only	Image data + Clinical data
Accuracy (%)	89.47	92.98
Sensitivity (%)	86.36	93.28
Specificity (%)	93.75	92.55
Positive predictive value (%)	95.00	94.69
Negative predictive value (%)	83.33	90.62
AUC	0.9751	0.9841

* Descriptive values are presented as percentages or actual values

AUC, area under the curve

correctly classified 114 cases as requiring surgery and 90 cases as eligible for conservative treatment, with 18 and 6 misclassified cases in each category, respectively. In contrast, the image and clinical data-integrating model correctly classified 125 cases as requiring surgery and 87 cases as eligible for conservative treatment, with only 9 and 7 misclassified cases, respectively. Table 6 shows the diagnostic performances metrics of the ML models. Although the integrating model showed slightly better accuracy and sensitivity, the McNemar test revealed no statistically significant difference in the classification performance between the models ($P = 0.573$). Figure 4 illustrates the ROC curves of both ML models with a 95%

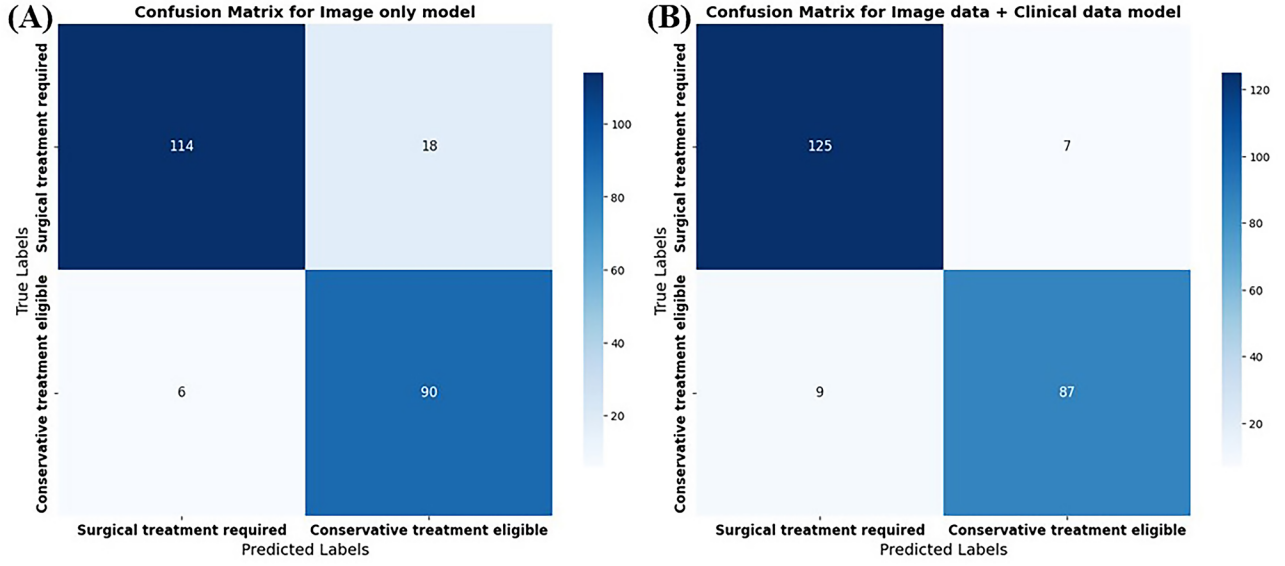

Fig. 3 **A.** Confusion matrix of the image-only model. **B.** Confusion matrix of the image and clinical data integrating model

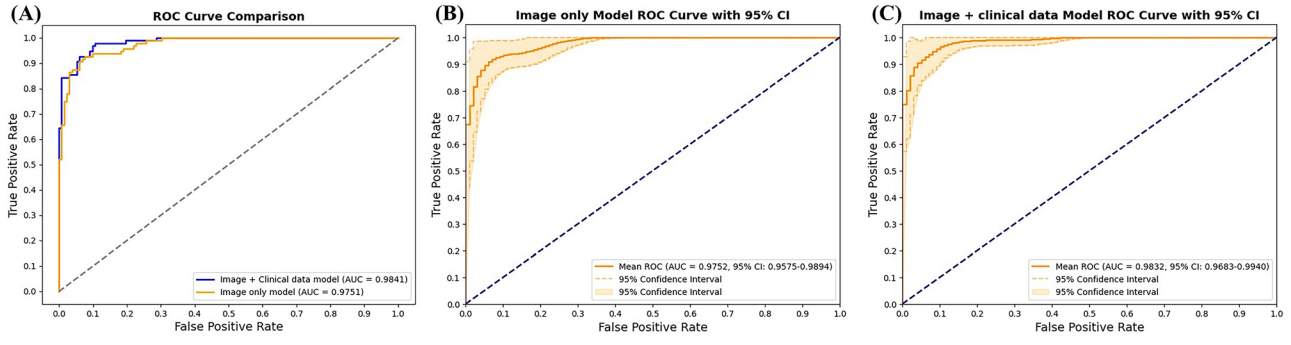

Fig. 4 Receiver operating characteristic (ROC) curve of the models (image-only model and image and clinical data integrating model). **A**. Comparison of ROC curves between the two models. **B**. ROC curve for the image-only model, including the 95% confidence interval (CI). **C**. ROC curve for the image and clinical data integrating model, including the 95% CI

confidence interval. The integrated model showed a marginal increase in the AUC compared to the image-only model; however, DeLong's test did not reveal a statistically significant difference ($P = 0.310$).

Figure 5 presents the results of the Grad-CAM analysis in the image and clinical data-integrating ML model. The Grad-CAM heatmaps exhibited that the radiocarpal joint, volar, and dorsal cortex of the radial metaphysis were highlighted regions for model training. Supplementary Fig. 2 presents the classification performance of the models that were retrained using the masking-based training strategy, with different threshold values applied to the Grad-CAM heatmaps. As the threshold decreased and a greater portion of the Grad-CAM-highlighted regions were masked, the model performance deteriorated accordingly. These results quantitatively demonstrate that the regions highlighted by the Grad-CAM not only serve as a visual explanation, as shown in Fig. 5, but also contribute substantially to the decision-making process of the model. Figure 6 displays the results of the SHAP analysis in the image and clinical data integrating ML model. The results indicated that female gender and the presence of subsequent or concomitant fractures were more strongly associated with the need for surgical treatment, whereas older age and lower BMI were weakly associated with the need for surgical treatment.

Discussion

Surgical decision-making is a complex process influenced by various factors, including the emotions and values of patients and caregivers, medical resource availability, rapport between patients and surgeons, time constraints, and individual judgments [33]. Surgical decision-making includes determining the appropriate surgical method and extent, deciding the time for surgery, and determining whether or not to perform surgery [34]. Determining whether to perform surgical treatment is a pivotal aspect of the surgical decision-making process. In this study, a ML model that can determine the need for surgical treatment in patients with DRFs was trained using a dataset that combines image data and clinical information. Accordingly, this study aimed to provide appropriate evidence for evaluating the need for surgical treatment and suggest the capability of the ML model as a decision-support tool.

The aforementioned results indicated that the accuracy of the two trained models was approximately 90%, and the image and clinical data-incorporating model had approximately 3% higher accuracy than the image-only

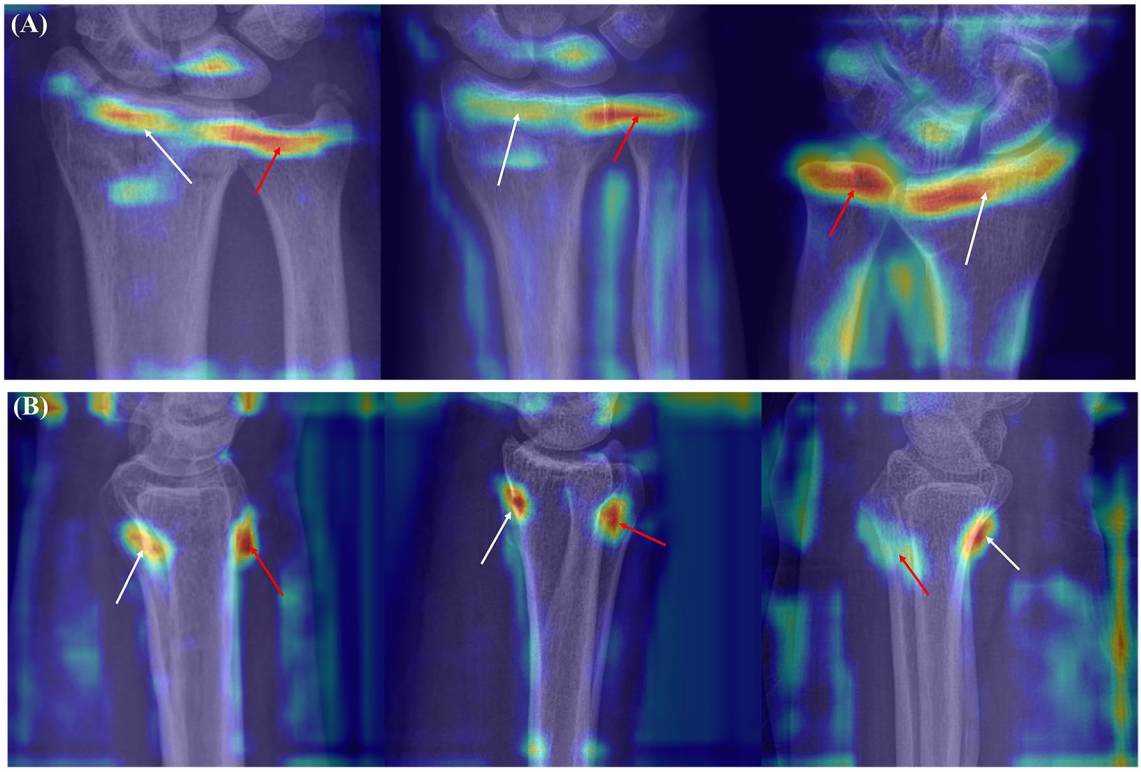

Fig. 5 **A.** Gradient-weighted Class Activation Mapping (Grad-CAM) visualization on anteroposterior wrist radiograph images. The heatmap indicates that the congruency of the radiocarpal joint (white arrow) was a critical region for model training. **B.** Grad-CAM visualization on lateral wrist radiograph images. The heatmap shows that the volar cortical disruption of the radial metaphysis (white arrow) and dorsal metaphyseal comminution (red arrow) were critical regions for model training

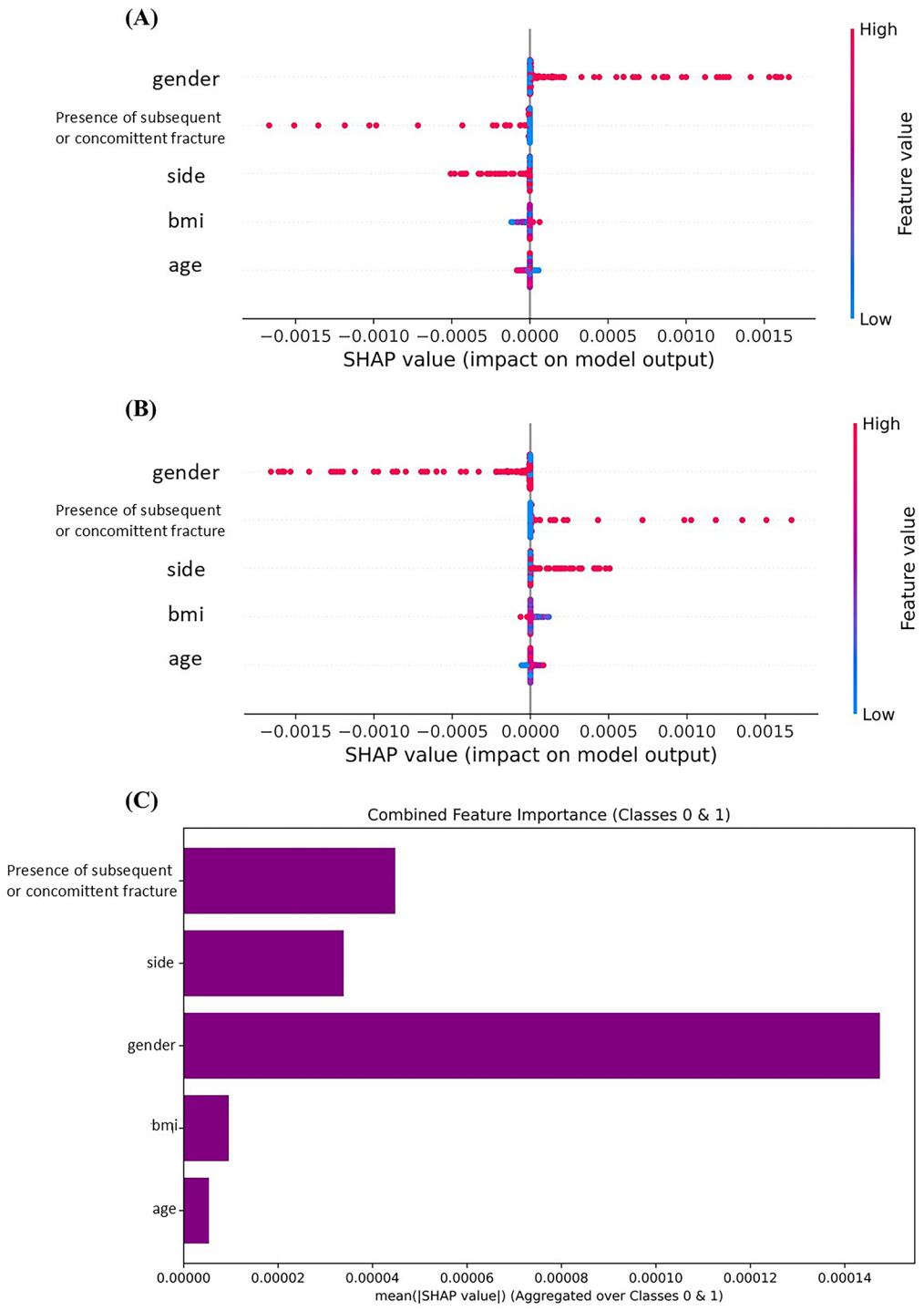

Fig. 6 **A.** SHapley Additive exPlanations (SHAP) values for clinical features in predicting the need for surgery (surgery required). Each dot represents a SHAP value for a single patient's feature, with its position indicating the degree of influence on the model's decision and its color representing the feature's value (red = high, blue = low). **B.** SHAP values for clinical features in predicting the need for surgery (surgery not required), following the same format as Fig. 6A, illustrating how these features influenced cases where surgery was not required. **C.** Mean absolute SHAP values across both classes, offering a global perspective on feature importance in the model's decision-making

model. This yield is comparable to, or slightly lower than, the performance of previous deep-learning models that focus exclusively on detecting fractures or classifying fracture types [35, 36, 37, 38]. Decisions regarding surgical treatment for fractures can be influenced by the type and severity of fractures observed in the radiographic images and factors such as the patient's age, gender, and underlying conditions [39, 40]. Moreover, the surgeon's preference and availability of hospital resources can also influence the decision-making process [33]. Therefore, surgical treatment decisions may be influenced by different clinical circumstances and subjective factors that were not included in the current model training. Consequently, model training using limited data could lead to inherent accuracy limitations. Nevertheless, the accuracy of approximately 90% of both models might be sufficient for diagnostic performance. Moreover, the better performance of the image and clinical data-integrating model compared with the image-only model indicates that incorporating clinical data in model training could play a crucial role in improving model's performance.

Although the difference in the AUC did not reach statistical significance based on DeLong's test, the observed improvement in AUC may still be considered clinically meaningful. The integrated model consistently yielded slightly superior results across multiple evaluation measures. This finding may be important from a clinical perspective, even if not statistically significant. In clinical situations such as surgical decision-making, which must carefully consider individualized patient factors, even small differences in predictive accuracy may influence treatment strategies. Furthermore, while the absolute increase in AUC may appear numerically small, this improvement can be regarded as meaningful when considering the dimensional disparity between the image and clinical data. The image data consisted of high-dimensional pixel-level features ($512 \times 512 \times 3$) from AP and lateral radiographic views. In contrast, the clinical data included only a few low-dimensional scalar variables such as age, gender, injured side, BMI, and the presence of concomitant or subsequent fractures. Despite the overwhelming volume and richness of the image features, the integrated model demonstrated further performance improvement over the image-only model by incorporating a small set of low-dimensional clinical features. This indicates that the addition of clinical information provided a complementary value that was not adequately represented by the image-only model. Although the improvement may not reach statistical significance according to standard statistical criteria, it reflects the benefit of multimodal integration, particularly in clinical settings where context-specific variables are critical in guiding surgical decision-making.

The results of the Grad-CAM analysis showed that the radiocarpal joint on AP radiographs and the volar and dorsal cortex of the radial metaphysis on lateral radiographs were important regions in determining the need for surgical treatment. A previous retrospective study indicated that surgical treatment for intra-articular DRFs may achieve improved functional outcomes [41]. Another study evaluated the risk factors associated with re-displacement following DRF reduction, identifying the initial degree of displacement and advanced age as significant risk factors [42]. In addition, in displaced DRFs, surgical treatment is more effective than nonsurgical treatment in improving radiographical and functional outcomes [43]. Therefore, based on a synthesis of the results of previous studies, the results of the Grad-CAM analysis in the present study imply that anatomical reduction followed by internal fixation may be necessary in cases of radiocarpal joint incongruency, volar cortical disruption of the distal radial metaphysis, and dorsal metaphyseal comminution. This might be considered joint incongruency, which significantly increases the risk of traumatic arthritis. In addition, the volar cortical disruption in the radial metaphysis and dorsal metaphyseal comminution could contribute to the progression of fracture displacement.

In this study, the SHAP results indicated that the female gender was most strongly associated with the need for surgical treatment, followed by the presence of subsequent or concomitant fractures. Compared with women, men typically have a larger skeletal structure and greater bone mass, which contributes to a lower incidence of stress and osteoporotic fractures [44]. In this study, the distribution of clinical data revealed the higher prevalence of fractures among women than among men, and the proportion of women was also higher in the group requiring surgical intervention. Moreover, the conservative treatment of concomitant fractures can disrupt early mobilization and delay rehabilitation associated with recovery [45]. These factors might influence the feature importance scores in the SHAP analysis.

Another notable result from the SHAP analysis was that left-side fractures were more likely to require surgical treatment than right-side fractures. This may be due to the equal distribution of left- and right-side fractures in the non-surgical group, whereas the surgical group had a slightly higher proportion of left-side fractures. However, given the lack of significant differences in the left- and right-side fracture distributions between the two groups, establishing its clinical relevance is challenging. The SHAP analysis also indicated a weak association among older age, lower BMI, and the need for surgical treatment. This might be associated with aging and lower BMI values, which contribute to the reduction in bone mineral density [46, 47]. However, several studies have

reported varying conclusions regarding the influence of age and BMI on increasing fracture severity and the need for surgical treatment [48, 49, 50, 51, 52]. They have included patient populations with diverse characteristics such as race, age, and gender, employed a retrospective design and analyzed various types of fractures. These factors may have contributed to the variability in their findings. Therefore, the SHAP analysis results may need to be carefully interpreted, and ongoing model training with a larger clinical dataset may be necessary.

The integrated ML model trained in this study identified several features that might be associated with surgical decision-making, such as female gender and the presence of concomitant or subsequent fractures; however, such features are not currently addressed in clinical practice guidelines for the management of DRFs [1, 43]. This discrepancy indicates a potential limitation in the current evidence-based guidelines, which may not yet incorporate up-to-date insights obtained through ML models. As the clinical utility of ML models continues to be validated through high-quality studies, future updates to treatment guidelines may be enhanced by integrating these predictive features to better support surgical decision-making.

In a previous study, a ML model was developed to determine the need for surgical treatment in patients with skeletal malocclusion using simple radiographs [53]. In that study, supervised ML was conducted using only radiographic data from two views, without concatenating clinical information in the analysis. However, determining the need for surgically treating DRFs based solely on radiographs is insufficient. The patient's age, presence of underlying conditions, desire for surgery, and adherence to postoperative rehabilitation are also considered important decision-making factors [39, 40]. Therefore, combining image data and clinical situation is necessary when making decisions in diverse clinical decision-making processes, such as determining surgical treatment. To address the limitations of the image-only model, clinical data were integrated into model training, leading to an improvement in model performance.

This study has several limitations. First, the plain radiographs and clinical data analyzed were sourced from an Asian population who attended a single hospital. Therefore, future studies involving diverse ethnic groups across multiple institutions are needed to improve generalizability. Second, the data set of 1,139 patients and 2,278 plain radiographs was considered small. To address this issue, image data augmentations were employed. Third, the developed ML model can only be applied to adult patients with DRFs. Growth plates in pediatric wrist radiographs may resemble fracture lines. Therefore, additional patient data collection and model training to differentiate between growth plates and fractures are required. Fourth, the clinical information used for model training was limited to certain factors, such as age, gender, and BMI. Considering that determining the need for surgical treatment depends on numerous clinical factors [33], including surgeon-related aspects, the clinical significance of the developed model should be carefully interpreted. Fifth, the data used for ML model training were obtained from a single institution, reflecting the specific preferences and clinical practices of that institution and its surgeons. Accordingly, the model's accuracy may be limited to that specific setting and may not adequately represent diverse ethnicities or geographic regions.

Conclusions

The ML model developed in this study may assist in assessing the necessity of surgical treatment for patients with DRFs. This model can be utilized to enhance the success rate of fracture treatment, support decision-making, and improve efficiency in the clinical setting. Given the high prevalence of DRFs and the increasing number of affected patients, ongoing efforts are necessary to improve the model's accuracy by increasing the number of training cases and incorporating up-to-date algorithms. Moreover, external validation of the proposed ML model using an independent dataset will be essential to evaluate its generalizability and robustness across different clinical settings.

Abbreviations

DRF	Distal radius fracture
ML	Machine learning
AP	Anteroposterior
EMR	Electric medial record
BMI	Body mass index
ICC	Intra-class correlation coefficient
ReLU	Rectified linear unit
Grad-CAM	Gradient-weighted Class Activation Mapping
SHAP	SHapley Additive exPlanations
AUC	Area under the curve
PPV	Positive predictive value
NPV	Negative predictive value

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13018-025-05830-z.

Supplementary Material 1: Figure 1. Detailed architecture of the U-Net feature extractor employed in this study. Each box represents the spatial resolution and number of channels of the feature map (height × width × channels) at each stage of the encoder and decoder. The central bottleneck layer compresses the high-level features prior to upsampling, and the skip connections between the corresponding encoder and decoder blocks are illustrated as horizontal arrows

Supplementary Material 1: Figure 2. Classification accuracy at varying thresholds using a masking-based training strategy

Acknowledgements

None declared.

Author contributions
Jongmin Lim (JL) and Sehun Chang (SC) contributed equally to this work and share co-first authorship. SWH, JL and SC: designed the study. SWH, HJP and EK: acquisition of data. All authors: analysis and interpretation of data. SWH, JL and SC: prepared and edited manuscript. All authors: read and approved the final manuscript.

Funding
This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00165960) and the Korea Internet & Security Agency (KISA) grant funded by the Korea government (PIPC) (No. RS-2023-00231200). The funders had no role in the study design, data collection, analysis, interpretation, and writing the manuscript.

Data availability
The datasets analyzed during the current study are not publicly available due to institutional policies and ethical considerations involving patient confidentiality, but they may be available from the corresponding author upon reasonable request and subject to institutional review board approval.

Declarations

Ethical approval and consent to participate
This study protocol was approved by the Institutional Review Board of university hospital (IRB No. KBSMC 2023-01-044) and the requirement was waived to obtain informed consent.

Competing interests
The authors declare no competing interests.

Received: 4 March 2025 / Accepted: 18 April 2025
Published online: 26 April 2025

References

1. Shapiro LM, Kamal RN, Kamal R, et al. Distal radius fracture clinical practice Guidelines–Updates and clinical implications. J Hand Surg. 2021;46(9):807–11.
2. Porrino JA Jr., Maloney E, Scherer K, Mulcahy H, Ha AS, Allan C. Fracture of the distal radius: epidemiology and premanagement radiographic characterization. AJR Am J Roentgenol. 2014;203(3):551–9.
3. Mauck BM, Swigler CW. Evidence-Based review of distal radius fractures. Orthop Clin North Am. 2018;49(2):211–22.
4. Alluri RK, Hill JR, Ghiassi A. Distal radius fractures: approaches, indications, and techniques. J Hand Surg Am. 2016;41(8):845–54.
5. Padegimas EM, Osei DA. Evaluation and treatment of osetoporotic distal radius fracture in the elderly patient. Curr Rev Musculoskelet Med. 2013;6(1):41–6.
6. Alzubaidi L, Al-Dulaimi K, Salhi A, et al. Comprehensive review of deep learning in orthopaedics: applications, challenges, trustworthiness, and fusion. Artif Intell Med. 2024;155:102935.
7. Alsoof D, McDonald CL, Kuris EO, Daniels AH. Machine learning for the orthopaedic surgeon: uses and limitations. J Bone Joint Surg Am. 2022;104(17):1586–94.
8. Kalmet PHS, Sanduleanu S, Primakov S, et al. Deep learning in fracture detection: a narrative review. Acta Orthop. 2020;91(2):215–20.
9. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ. Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. J Digit Imaging. 2020;33(5):1209–17.
10. van der Gaast N, Bagave P, Assink N, et al. Deep learning for tibial plateau fracture detection and classification. Knee. 2025;54:81–9.
11. Min H, Rabi Y, Wadhawan A, et al. Automatic classification of distal radius fracture using a two-stage ensemble deep learning framework. Phys Eng Sci Med. 2023;46(2):877–86.
12. McMaster C, Bird A, Liew DFL, et al. Artificial intelligence and deep learning for rheumatologists. Arthritis Rheumatol. 2022;74(12):1893–905.
13. Dong Q, Luo G, Lane NE, et al. Deep learning classification of spinal osteoporotic compression fractures on radiographs using an adaptation of the Genant semiquantitative criteria. Acad Radiol. 2022;29(12):1819–32.
14. Hong N, Cho SW, Shin S, et al. Deep-Learning-Based detection of vertebral fracture and osteoporosis using lateral spine X-Ray radiography. J Bone Min Res. 2023;38(6):887–95.
15. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell. 2020;2(2):e190023.
16. Magneli M, Ling P, Gislen J, et al. Deep learning classification of shoulder fractures on plain radiographs of the humerus, scapula and clavicle. PLoS ONE. 2023;18(8):e0289808.
17. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 2018;89(4):468–73.
18. Yoon AP, Lee YL, Kane RL, Kuo CF, Lin C, Chung KC. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open. 2021;4(5):e216096.
19. Oude Nijhuis KD, Dankelman LHM, Wiersma JP, et al. AI for detection, classification and prediction of loss of alignment of distal radius fractures; a systematic review. Eur J Trauma Emerg Surg. 2024;50(6):2819–31.
20. Hornung AL, Rudisill SS, Smith S, Streepy JT, Simcock XC. Can machine learning identify patients who are appropriate for outpatient open reduction and internal fixation of distal radius fractures?? J Hand Surg Glob Online. 2024;6(6):808–13.
21. Breu R, Avelar C, Bertalan Z, et al. Artificial intelligence in traumatology. Bone Joint Res. 2024;13(10):588–95.
22. Holzinger A, Haibe-Kains B, Jurisica I. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. Eur J Nucl Med Mol Imaging. 2019;46(13):2722–30.
23. Jo H, Kim C, Gwon D, et al. Combining clinical and imaging data for predicting functional outcomes after acute ischemic stroke: an automated machine learning approach. Sci Rep. 2023;13(1):16926.
24. Lin CY, Guo SM, Lien JJ, et al. Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT. Radiol Med. 2024;129(1):56–69.
25. Benjamins JW, Yeung MW, Maaniitty T, et al. Improving patient identification for advanced cardiac imaging through machine learning-integration of clinical and coronary CT angiography data. Int J Cardiol. 2021;335:130–6.
26. Chlap P, Min H, Vandenberg N, Dowling J, Holloway L, Haworth A. A review of medical image data augmentation techniques for deep learning applications. J Med Imaging Radiat Oncol. 2021;65(5):545–63.
27. Demircioglu A. The effect of feature normalization methods in radiomics. Insights Imaging. 2024;15(1):2.
28. Azad R, Aghdam EK, Rauland A, et al. Medical image segmentation review: the success of U-Net. IEEE Trans Pattern Anal Mach Intell. 2024;46(12):10076–95.
29. Zhang X, Guo E, Liu X, et al. Enhancing furcation involvement classification on panoramic radiographs with vision Transformers. BMC Oral Health. 2025;25(1):153.
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via Gradient-Based localization. Int J Comput Vision. 2020;128(2):336–59.
31. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Paper presented at: Neural Information Processing Systems2017.
32. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.
33. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical Decision-making. JAMA Surg. 2020;155(2):148–58.
34. Loftus TJ, Filiberto AC, Li Y, et al. Decision analysis and reinforcement learning in surgical decision-making. Surgery. 2020;168(2):253–66.
35. Anttila TT, Karjalainen TV, Makela TO, et al. Detecting distal radius fractures using a Segmentation-Based deep learning model. J Digit Imaging. 2023;36(2):679–87.
36. Oka K, Shiode R, Yoshii Y, Tanaka H, Iwahashi T, Murase T. Artificial intelligence to diagnosis distal radius fracture using biplane plain X-rays. J Orthop Surg Res. 2021;16(1):694.
37. Raisuddin AM, Vaattovaara E, Nevalainen M, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep. 2021;11(1):6006.
38. Tobler P, Cyriac J, Kovacs BK, et al. AI-based detection and classification of distal radius fractures using low-effort data labeling: evaluation of applicability and effect of training set size. Eur Radiol. 2021;31(9):6816–24.
39. Szatmary P, Arora S, Sevdalis N. To operate or not to operate? A multimethod analysis of decision-making in emergency surgery. Am J Surg. 2010;200(2):298–304.

40. Chhabra KR, Sacks GD, Dimick JB. Surgical decision making: challenging dogma and incorporating patient preferences. JAMA. 2017;317(4):357–8.
41. Sharma H, Khare GN, Singh S, Ramaswamy AG, Kumaraswamy V, Singh AK. Outcomes and complications of fractures of distal radius (AO type B and C): volar plating versus nonoperative treatment. J Orthop Sci. 2014;19(4):537–44.
42. Jung HW, Hong H, Jung HJ, et al. Redisplacement of distal radius fracture after initial closed reduction: analysis of prognostic factors. Clin Orthop Surg. 2015;7(3):377–82.
43. Kamal RN, Shapiro LM. American academy of orthopaedic surgeons/american society for surgery of the hand clinical practice guideline summary management of distal radius fractures. J Am Acad Orthop Surg. 2022;30(4):e480–6.
44. Nieves JW, Formica C, Ruffing J, et al. Males have larger skeletal size and bone mass than females, despite comparable body size. J Bone Min Res. 2005;20(3):529–35.
45. Kang SW, Shin WC, Moon NH, Suh KT. Concomitant hip and upper extremity fracture in elderly patients: prevalence and clinical implications. Injury. 2019;50(11):2045–8.
46. Hsu S, Bansal N, Denburg M, et al. Risk factors for hip and vertebral fractures in chronic kidney disease: the CRIC study. J Bone Min Res. 2024;39(4):433–42.
47. de Melo TG, da Assumpcao LV, Santos Ade O, Zantut-Wittmann DE. Low BMI and low TSH value as risk factors related to lower bone mineral density in postmenospausal women under Levothyroxine therapy for differentiated thyroid carcinoma. Thyroid Res. 2015;8:7.
48. Kanis JA, Johnell O, Oden A, Dawson A, De Laet C, Jonsson B. Ten year probabilities of osteoporotic fractures according to BMD and diagnostic thresholds. Osteoporos Int. 2001;12(12):989–95.
49. Sander AL, Leiblein M, Sommer K, Marzi I, Schneidmuller D, Frank J. Epidemiology and treatment of distal radius fractures: current concept based on fracture severity and not on age. Eur J Trauma Emerg Surg. 2020;46(3):585–90.
50. Kloberdanz AL, Meyer J, Kammermeier K, et al. Impact of body mass index on fracture severity, clinical, radiological and functional outcome in distal radius fractures: a retrospective observational study after surgical treatment. Arch Orthop Trauma Surg. 2024;144(6):2915–23.
51. Kim SH, Yi SW, Yi JJ, Kim YM, Won YJ. Association between body mass index and the risk of hip fracture by sex and age: A prospective cohort study. J Bone Min Res. 2018;33(9):1603–11.
52. Cui P, Wang W, Wang Z, et al. The association between body mass index and bone mineral density in older adults: a cross-sectional study of community population in Beijing. BMC Musculoskelet Disord. 2024;25(1):655.
53. Shin W, Yeom HG, Lee GH, et al. Deep learning based prediction of necessity for orthognathic surgery of skeletal malocclusion using cephalogram in Korean individuals. BMC Oral Health. 2021;21(1):130.

Publisher's note