Research Square

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

The Efficacy of Artificial Intelligence Models in Diagnosing Knee Fractures from X-ray Images using Hybrid Attention Architecture

Francisco Fernández Schlein
Pontificia Universidad Católica de Chile
Rodrigo De Marinis Acle
ridemarinis@uc.cl
Pontificia Universidad Católica de Chile
Luis Irribarra Trivelli
Pontificia Universidad Católica de Chile
Mario Orrego Luzoro
Pontificia Universidad Católica de Chile

Research Article

Keywords: Knee fractures, Artificial Intelligence, Deep Learning, Convolutional Neural Networks, Transfer Learning

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8535392/v1

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

Additional Declarations: No competing interests reported.

Abstract

Purpose: This study assesses the efficacy of ten artificial intelligence (AI) models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) models, for detecting knee fractures using a multi-view attention mechanism to integrate anteroposterior (AP) and lateral (LAT) X-rays.

Methods: A retrospective dataset of 860 radiographs sets (430 fractures, 430 controls) was used to train ten deep learning models. The models employed transfer learning for feature extraction and a custom attention-based block for multi-view fusion. Data augmentation was employed to address limited sample sizes and overfitting risks. Performance was evaluated using sensitivity, specificity, F1-score, and area under the curve (AUC).

Results: ViT Small, MobileNetV3, and EfficientNetV2 achieved the highest AUC (0.94). Median sensitivity, specificity, F1-score, and AUC across all models were 0.87, 0.80, 0.86, and 0.93, respectively, with no statistically significant differences observed ($p > 0.05$). The median image processing time was 68.57 milliseconds per image, with EfficientNetV2 being the fastest.

Conclusion: The AI models tested demonstrate high diagnostic accuracy for knee fractures using a multi-view attention approach, performing effectively on small-to-medium datasets independently of the specific architecture (CNN or ViT). Their rapid processing speed and adaptability highlight their potential utility as clinical decision support tools, particularly in limited-resource settings.

Keywords: Knee fractures, Artificial Intelligence, Deep Learning, Convolutional Neural Networks, Transfer Learning.

1. Introduction

Fractures around the knee account for 3.9% of global cases, ranging from 733 to 4017 per 100,000 [1–4]. Surgical management is generally preferred over non-operative treatment due to higher rates of post-traumatic osteoarthritis and poorer clinical outcomes. [5]. These fractures are costly, impacting healthcare systems and causing significant work absenteeism. Timely diagnosis and appropriate management are critical to minimizing complications and improving outcomes [6].

Plain radiographs remain the first-line imaging modality for diagnosing fractures around the knee [7, 8]. However, their sensitivity is limited, particularly for fractures involving the articular surface [9]. Therefore, it is essential to set a low threshold for using advanced studies like computed tomography (CT) scans to confirm the diagnosis when an initial suspicion exists [9, 10]. Initial assessments are frequently performed in emergency departments (ED) by non-specialist physicians, which contributes to a reported 3.1% missed diagnosis rate [11].

In recent years, computer-aided diagnostic (CAD) tools have gained attention for improving musculoskeletal imaging interpretation [12–14]. Among these, artificial intelligence (AI) models — particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) — have demonstrated state-of-the-art performance in image analysis tasks, due to their performance and growing accessibility [15, 16].

These models share some capabilities regarding image analysis, but have some structural differences in how they interpret and process the input data. CNNs are highly effective at obtaining local features through their multiple convolutional filters but struggle to capture context and long-range associations. This limitation is addressed by ViT models, which employ image patch decomposition and positional embedding, similar to Large Language Models, to incorporate contextual information and model long-distance relationships during inference [15].

Evidence on AI performance for knee fracture detection using X-rays is scarce. Lind et al. reported a mean area under the curve (AUC) of 0.88 with a ResNet-based model for AO/OTA classification [17], while Cohen et al. achieved an AUC near 0.80 when detecting indirect signs such as effusion, though these

findings lack specificity [18]. A recent meta-analysis of AI-assisted fracture detection showed pooled sensitivity and specificity of 0.90 and 0.92, comparable to expert interpretation [19]. These results are promising, but most studies remain generalized and not specifically focused on knee fractures.

This study evaluates ten deep learning models (CNNs and ViTs) for knee fracture diagnosis, utilizing a multi-view attention mechanism to mimic standard clinical assessment. We hypothesize that integrating anteroposterior and lateral views will yield high diagnostic accuracy even with limited training data, validating AI as a practical decision-support tool for orthopedic trauma.

2. **Material and Methods**

This work follows the recommendations of the CLAIM guideline for reporting AI works related to medical imaging [20] (Supplementary data S1) and it was approved by our institutional review board (IRB no. 230203002).

2.1 **Dataset**

We retrospectively reviewed an institutional database of knee radiographs studies interpreted by expert's radiologist, in the period between 2015 – 2022 diagnosed with any fracture. We include only images with closed physis and complete examinations, including at least anteroposterior (AP) and lateral (LAT) views. Exclusion criteria included the presence of osteosynthesis hardware, advanced osteoarthritis (Kellgren–Lawrence grade II or higher) [21], duplicate studies and images with cast or immobilization devices. After filtering, 430 fracture cases were identified and stratified paired by age and sex with 430 controls, yielding a total of 860 radiograph sets (Fig 1).

2.2 **Preprocessing**

The dataset was split into training (80%), validation (10%), and testing (10%) subsets, stratified by class labels to maintain fracture prevalence. All DICOM images were converted to PNG format and resized to 224 × 224 pixels (three channels) to ensure compatibility with pre-trained models. Pixel values were normalized using ImageNet mean and standard deviation statistics [22]. For ViT models, images were divided into fixed-size patches, depending on the specific model, which were then flattened and linearly projected into feature embeddings. In contrast, CNN models processed the full pixel grid directly.

To address the limited dataset size and prevent overfitting, we implemented a rigorous data augmentation pipeline, including geometrical and light contrast transformations such as horizontal and vertical flips, shearing, cropping, and random brightness adjustments. This approach generates transformed images for each batch during every iteration, effectively creating new data from existing samples without increasing the total number of images in the dataset [23].

2.3 **Model Architecture**

Fracture detection was approached as a binary classification task. Deep learning models, specifically Convolutional Neural Networks (CNN) and Vision Transformers (ViT), were employed as feature extractors. CNNs, such as ResNet [24] and DenseNet [25], uses multiple convolutional layers to effectively capture textures and edges in images [26, 27]. In contrast, ViTs process images as sequences of patches using self-attention mechanisms, allowing them to capture global context and long-range dependencies across the entire image [15]. To utilize the complementary information provided by standard radiographic series, we designed a "Dual-Stream" architecture consisting of three distinct blocks: feature extraction (FE), multi-view fusion, and classification (Fig 2).

Feature extraction

The process begins with an initial FE block with two independent-separate parallel backbones (one for AP and one for LAT projections). This independence allows each branch to learn and capture specific view-features. We used the same model in each stream as FE in every test, which could be either a CNN or a ViT model, selected from a pre-defined list of the most frequently cited architectures in the literature (Table 3) [28]. We did not combine different backbones to reduce variability given the large number of potential combinations. For CNN backbones, the spatial feature maps output by the final convolutional layer are flattened into a sequence of vectors, while for ViT backbones, the patch embeddings are used directly. This tokenization process standardizes the output from both architecture types into a sequence format compatible with the subsequent attention mechanism (Fig 2).

Multi-view fusion

Following the FE block, each feature sequence from AP and LAT are lineally concatenated obtaining one single joint representation. Next, we applied a Multi-Head Self-Attention (MHA) mechanism with four parallel heads to integrate the information extracted from each view. Unlike simple vector concatenation, this approach enables each projection to attend to the other, implementing an “all-to-all” communication strategy [29]. This design allows the model to dynamically learn relationships between different regions of the knee from different views and integrate them. The output of the attention layer is a sequence of context-enriched vectors, which is then passed through a Global Average Pooling layer to condense the sequence into a single feature vector. This vector is subsequently forwarded to the final classification layer.

Classification

The final stage is the classification block. The fused pooling feature vector is passed through a Multi-Layer Perceptron (MLP) consisting of a simple linear layer, a ReLU activation function, and a dropout layer (rate = 0.2) to further regularize the model [30]. The final linear projection produces a scalar logit, which is converted into a fracture probability using a sigmoid function, followed by a simple decision threshold set at 0.5.

2.4 **Model training, implementation and assessment.**

The models were trained using PyTorch, a dedicated AI python library, using the framework PyTorch Lightning for simplicity and user-friendly handling. Model development, training, and testing were conducted on a cloud-based platform (Lightning AI®), using NVIDIA T4 GPUs with 16 GB of VRAM, featuring Tensor Cores optimized for deep learning tasks. Training was performed over 50 epochs using the Binary Cross-Entropy with Logits loss function, an adaptive learning rate schedule, and early stopping to prevent overfitting (see Table 1). The fusion and classification stages were trained from randomly initialized weights, while the feature extraction block was fine-tuned using pre-trained weights from ImageNet, and additionally PubMed-CLIP weights for the BioMedCLIP ViT model [31, 32].

Attention maps were generated using the Grad-Cam algorithm, which highlights the regions of an image most influential for a given class prediction. This is achieved by analyzing the gradients in the last convolutional layer and computing the predicted class score for each feature map for a specific class. The

result is a heatmap that highlights the crucial regions of the image, providing insights into the network's decision-making process [33].

Continuous and categorical variables were described as means (SD) and frequencies (%), respectively. Model performance (Sensitivity, Specificity, F1-score [Eq 1], AUC) was reported with 95% confidence intervals estimated by bootstrapping (1,000 iterations). Statistical heterogeneity across models was assessed using Cochran's Q test, while median performance differences between model families (ViT vs. CNN) were evaluated with the Mann-Whitney U test, and AUC comparisons with the DeLong test. Statistical significance was defined as a p-value < 0.05 and analyses were performed using Stata 14.0 and Python (v3.12) libraries.

$$F1 = 2 * \left(\frac{Precision * Sensitivy}{Precision + Sensitivy}\right), \quad (1)$$

3. **Results**

The dataset comprised 430 fractures and 430 controls. The fracture group had a mean age of 56.7 years (SD ± 20.3) with 52.5% women, while the control group had a mean age of 56.4 years (SD ± 18.8) with 58.3% women. No significant differences in gender distribution were observed (p = 0.9). In both groups, females were significantly older than males (p < 0.05). Fracture distribution was predominantly tibial (47.4%) and patellar (32.6%). Detailed demographics are provided in Table 2.

Performance metrics were computed for all ten models (Table 3). ViT-Small achieved the highest sensitivity (0.92; 95% CI: 0.86–0.97), followed closely by MobileNetV3, Swin-Tiny, and DenseNet121 (all 0.91). Regarding specificity, ConvNeXt Tiny and DeiT-Small demonstrated the highest performance, both achieving 0.85 (95% CI: 0.76–0.93 and 0.77–0.93, respectively). The median sensitivity and specificity across all models were 0.87 and 0.80, respectively. Statistical analysis revealed significant heterogeneity in sensitivity across the models (p < 0.001); however, when comparing the median performance between model families (CNN vs. ViT), no statistically significant differences were found for sensitivity, specificity, or any other metric (p > 0.05).

In terms of AUC, ViT-Small, MobileNetV3, and EfficientNetV2 achieved the top performance, all scoring 0.94 (95% CI: 0.90–0.97 for ViT-Small/EfficientNetV2 and 0.89–0.97 for MobileNetV3). These were followed

by Swin-Tiny, BioMedCLIP, and ConvNeXt Tiny, which all reached an AUC of 0.93. No statistically significant differences were observed in AUC values among the ten models ($p > 0.05$), these are illustrated in Figure 3. Consistent with these findings, ViT-Small and MobileNetV3 also achieved the highest F1-score of 0.88, surpassing the group median of 0.86. Inference time was also assessed, yielding a median of 68.57 milliseconds (ms) per image. EfficientNetV2 was the fastest model (52.89 ms), whereas BioMedCLIP was the slowest (323.33 ms). Attention maps generated by Grad-CAM (Figure 4) illustrate representative cases, highlighting the model's ability to focus on fracture zones in true positives while ignoring irrelevant regions in negatives.

4. **Discussion**

In this study, ten AI models —spanning CNNs and ViTs— were trained to detect knee fractures using a multi view attention mechanism that integrates AP and LAT projections. Across architectures, diagnostic performance was high: ViT-Small, EfficientNetV2, and MobileNetV3 achieved the best area under the curve (AUC) of 0.94, with group medians of 0.87 for sensitivity, 0.80 for specificity, and 0.93 for AUC. Notably, no statistically significant performance differences were observed between CNN and ViT families ($p > 0.05$), indicating that with robust feature fusion, efficient CNNs can match the accuracy of more complex Transformer backbones. Inference was rapid (median 68.57 ms per image), with EfficientNetV2 being the fastest (52.89 ms), supporting the feasibility of deployment in clinical environments with standard hardware.

Published evidence on AI for knee fracture diagnosis remains limited but generally encouraging. Lind et al. trained a ResNet based model for AO/OTA classification and reported AUCs around 0.89, demonstrating viability for knee radiography but without multi view fusion [17]. More broadly, systematic reviews have shown pooled sensitivities of ~0.87–0.92 and specificities of ~0.90–0.91 across fracture detection tasks, with performance comparable to physicians, although methodological heterogeneity and risk of bias were frequently noted [34–36]. Our findings align with these trends: despite a moderate dataset ($n = 860$), dual view integration achieved state of the art accuracy within the reported ranges. Unlike studies that trained X-ray models using CT as ground truth—an approach that can be affected by known interobserver variability when mapping CT based classifications back to plain radiographs—our work adheres to a strictly radiographic ground truth, reflecting real world constraints in the ED pathway [37, 38].

A key innovation is the integration strategy. Previous orthopedic imaging studies have shown that multi view inputs outperform single view models for distal radius or hip fracture detection, often via feature concatenation across projections [39, 40]. However, simple stacking may not capture complex spatial dependencies between orthogonal views. We therefore adopted Multi Head Self Attention (MHA) to enable “all to all” cross view communication, allowing the model to dynamically weight AP versus LAT information per case [15, 29]. This design is clinically meaningful: when a fracture line is obscured in one plane but conspicuous in the other, attention can prioritize the informative features, contributing to the high AUC (0.94) and stability across backbones.

From a clinical implementation perspective, the absence of performance disparity between CNN and ViT families suggests that the backbone is less critical than the data integration strategy. Lightweight CNNs such as MobileNetV3 achieved top tier accuracy (AUC = 0.94) with shorter inference times (~53 ms) than ViT-Small (~68 ms), enabling practical deployment on standard workstations or even mobile devices in resource limited settings without compromising diagnostic performance. This efficiency is relevant for ED workflows in which throughput and latency directly influence patient triage and time to diagnosis.

To address interpretability—often cited as a limitation of deep learning—we generated Grad CAM saliency maps to visualize the regions driving model predictions [33]. In true positive cases, heatmaps consistently highlighted fracture zones, whereas true negative cases showed diffuse or absent focal activation, enhancing transparency. These visualizations help clinicians interpret model reasoning and reduce diagnostic uncertainty.

This study has limitations. Fracture heterogeneity (from subtle nondisplaced lines to complex comminution) imposes a challenging learning. The dataset, while balanced and stratified, was single institutional, which may limit generalizability. Despite augmentation and early stopping, external validation and prospective studies are needed to confirm performance and clinical impact. Future work should focus on external validation and evaluating calibration thresholds to optimize sensitivity versus specificity for different clinical triage scenarios, assess subgroup performance (e.g., age, sex, fracture location), and explore integration with clinical metadata. Finally, while attention mechanisms enhance fusion, comparative

studies against alternative cross view integration strategies and additional imaging modalities (e.g., CT when indicated) could clarify the boundaries of radiographic AI in complex knee trauma.

In summary, AI models that mimic clinical practice by jointly analyzing AP and LAT projections can achieve high diagnostic accuracy for knee fractures (AUC up to 0.94), with performance independent of backbone family when an effective fusion strategy is employed. The combination of accuracy, speed, and interpretability underscores the potential of these tools as accessible decision support systems in orthopedic trauma, particularly in settings where rapid, reliable X ray interpretation is essential.

5. **Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6. References

1. Bergh C, Wennergren D, Möller M, Brisby H (2020) Fracture incidence in adults in relation to age and gender: A study of 27,169 fractures in the Swedish Fracture Register in a well-defined catchment area. PloS One 15:e0244291. https://doi.org/10.1371/journal.pone.0244291

2. Curtis EM, van der Velde R, Moon RJ, et al (2016) Epidemiology of fractures in the United Kingdom 1988-2012: Variation with age, sex, geography, ethnicity and socioeconomic status. Bone 87:19–26. https://doi.org/10.1016/j.bone.2016.03.006

3. Vun S h., Aitken S a., McQueen M m., Court-Brown C m. (2013) Adult fracture patterns: epidemiology of fractures around the knee. Orthop Proc 95-B:10–10. https://doi.org/10.1302/1358-992X.95BSUPP_12.BOTA2012-010

4. Amin S, Achenbach SJ, Atkinson EJ, et al (2014) Trends in fracture incidence: a population-based study over 20 years. J Bone Miner Res Off J Am Soc Bone Miner Res 29:581–589. https://doi.org/10.1002/jbmr.2072

5. Wasserstein D, Henry P, Paterson JM, et al (2014) Risk of Total Knee Arthroplasty After Operatively Treated Tibial Plateau Fracture: A Matched-Population-Based Cohort Study. JBJS 96:144–150. https://doi.org/10.2106/JBJS.L.01691

6. Bonafede M, Espindle D, Bower AG (2013) The direct and indirect costs of long bone fractures in a working age US population. J Med Econ 16:169–178. https://doi.org/10.3111/13696998.2012.737391

7. Stiell IG, Wells GA, McDowell I, et al (1995) Use of radiography in acute knee injuries: need for clinical decision rules. Acad Emerg Med Off J Soc Acad Emerg Med 2:966–973. https://doi.org/10.1111/j.1553-2712.1995.tb03123.x

8. Jackson JL, O'Malley PG, Kroenke K (2003) Evaluation of Acute Knee Pain in Primary Care. Ann Intern Med 139:575. https://doi.org/10.7326/0003-4819-139-7-200310070-00010

9. Pinto A, Berritto D, Russo A, et al (2018) Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department. Acta Bio Medica Atenei Parm 89:111–123. https://doi.org/10.23750/abm.v89i1-S.7015

10. Mthethwa J, Chikate A (2018) A review of the management of tibial plateau fractures. Musculoskelet Surg 102:119–127. https://doi.org/10.1007/s12306-017-0514-8

11. Hallas P, Ellingsen T (2006) Errors in fracture diagnoses in the emergency department – characteristics of patients and diurnal variation. BMC Emerg Med 6:4. https://doi.org/10.1186/1471-227X-6-4

12. Chan H-P, Hadjiiski LM, Samala RK (2020) Computer-Aided Diagnosis in the Era of Deep Learning. Med Phys 47:e218–e227. https://doi.org/10.1002/mp.13764

13. Ko S, Pareek A, Ro DH, et al (2022) Artificial intelligence in orthopedics: three strategies for deep learning with orthopedic specific imaging. Knee Surg Sports Traumatol Arthrosc 30:758–761. https://doi.org/10.1007/s00167-021-06838-8

14. Chen K, Stotter C, Klestil T, Nehrer S (2022) Artificial Intelligence in Orthopedic Radiography Analysis: A Narrative Review. Diagnostics 12:2235. https://doi.org/10.3390/diagnostics12092235

15. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

16. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition

17. Lind A, Akbarian E, Olsson S, et al (2021) Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system. PLOS ONE 16:e0248809. https://doi.org/10.1371/journal.pone.0248809

18. Cohen I, Sorin V, Lekach R, et al (2024) Artificial intelligence for detection of effusion and lipo-hemarthrosis in X-rays and CT of the knee. Eur J Radiol 175:111460. https://doi.org/10.1016/j.ejrad.2024.111460

19. Zhang X, Yang Y, Shen Y-W, et al (2022) Diagnostic accuracy and potential covariates of artificial intelligence for diagnosing orthopedic fractures: a systematic literature review and meta-analysis. Eur Radiol 32:7196–7216. https://doi.org/10.1007/s00330-022-08956-4

20. Tejani AS, Klontzas ME, Gatti AA, et al (2024) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiol Artif Intell 6:e240300. https://doi.org/10.1148/ryai.240300

21. Kohn MD, Sassoon AA, Fernando ND (2016) Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. Clin Orthop 474:1886–1893. https://doi.org/10.1007/s11999-016-4732-4

22. Russakovsky O, Deng J, Su H, et al (2015) ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115:211–252. https://doi.org/10.1007/s11263-015-0816-y

23. Mikołajczyk A, Grochowski M (2018) Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). pp 117–122

24. He K, Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition

25. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2018) Densely Connected Convolutional Networks

26. O'Shea K, Nash R (2015) An Introduction to Convolutional Neural Networks

27. Alzubaidi L, Zhang J, Humaidi AJ, et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:53. https://doi.org/10.1186/s40537-021-00444-8

28. Su Z, Adam A, Nasrudin MF, et al (2023) Skeletal Fracture Detection with Deep Learning: A Comprehensive Review. Diagnostics 13:3245. https://doi.org/10.3390/diagnostics13203245

29. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is All you Need. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.

30. Srivastava N, Hinton G, Krizhevsky A, et al Dropout: A Simple Way to Prevent Neural Networks from Overfitting

31. ImageNet: A large-scale hierarchical image database | IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/5206848. Accessed 28 May 2024

32. Zhang S, Xu Y, Usuyama N, et al (2025) BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs

33. Selvaraju RR, Cogswell M, Das A, et al (2020) Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis 128:336–359. https://doi.org/10.1007/s11263-019-01228-7

34. Kuo RYL, Harrison C, Curran T-A, et al (2022) Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology 304:50–62. https://doi.org/10.1148/radiol.211785

35. Jung J, Dai J, Liu B, Wu Q (2024) Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. PLOS Digit Health 3:e0000438. https://doi.org/10.1371/journal.pdig.0000438

36. Yang S, Yin B, Cao W, et al (2020) Diagnostic accuracy of deep learning in orthopaedic fractures: a systematic review and meta-analysis. Clin Radiol 75:713.e17-713.e28. https://doi.org/10.1016/j.crad.2020.05.021

37. van der Gaast N, Bagave P, Assink N, et al (2025) Deep learning for tibial plateau fracture detection and classification. The Knee 54:81–89. https://doi.org/10.1016/j.knee.2025.02.001

38. Besa P, Angulo M, Vial R, et al (2023) The AO classification system for tibial plateau fractures: An independent inter and intraobserver agreement study. Injury 54:. https://doi.org/10.1016/j.injury.2023.04.028

39. Blüthgen C, Becker AS, Vittoria de Martini I, et al (2020) Detection and localization of distal radius fractures: Deep learning system versus radiologists. Eur J Radiol 126:108925. https://doi.org/10.1016/j.ejrad.2020.108925

40. Wang L-X, Zhu Z-H, Chen Q-C, et al (2024) Development and validation of a deep-learning model for the detection of non-displaced femoral neck fractures with anteroposterior and lateral hip radiographs. Quant Imaging Med Surg 14:527–539. https://doi.org/10.21037/qims-23-814

Figures and Tables

Table 1: Hyperparameter values for AI model training process. * Not all the models were trained over a fifty epochs; our strategy included an early stop to look at the validation loss value.

Hyperparameter	Value	Description
Optimizer	AdamW	Adaptive Moment Estimation with Weight Decay
Scheduler	Cosine Annealing	Warm Restarts ($T_0 = 5$, $T_{mult} = 1$)
Learning Rate	1×10^{-4}	Initial learning rate
Weight Decay	1×10^{-2}	Regularization factor
Batch Size	32	Images per training step
Epochs	50	Maximum training epochs
Early Stopping	Enabled	Monitor: Validation Loss
Attention Heads	4	Number of parallel attention heads, in MHA
Attn Dropout	0.1	Dropout probability inside MHA
Cls Dropout	0.2	Dropout probability in Classifier MLP

Table 2: Demographic description of the dataset. * Show the p-values related to the comparison between male and female for each group.

Characteristic	Fracture Group	Control Group	p-value
Total Cases (n)	430	430	-
Mean Age (years, SD)	56.7 (± 20.3)	56.4 (± 18.8)	-
Women (n, %)	226 (52.5%)	251 (58.3%)	0.9
Men (n, %)	204 (47.5%)	179 (41.7%)	0.9
Female Mean Age (years, SD)	64 (± 21.3)	62 (± 17.6)	< 0.05 *
Male Mean Age (years, SD)	49 (± 15.9)	49 (± 17-5)	< 0.05 *
Fracture Location	**Distribution**	-	-
Tibia	47.4%	-	-
Patella	32.55%	-	-
Femur	16.3%	-	-
Fibula	3.7%	-	-

Table 3: Comparative performance of Vision Transformer (ViT) and Convolutional Neural Network (CNN) models for knee fracture diagnosis.

Model Architecture	Model	AUC [a]	Sensitivity [b]	Specificity	F1-score	Time (ms)
ViT	ViT-Small	**0.94 [0.90, 0.97]**	**0.92 [0.86, 0.97]**	0.79 [0.69, 0.88]	0.88 [0.83, 0.93]	68.06
	Swin-Tiny	0.93 [0.89, 0.97]	0.91 [0.84, 0.96]	0.78 [0.68, 0.87]	0.87 [0.82, 0.92]	66.15
	BioMedCLIP	0.93 [0.89, 0.96]	0.86 [0.78, 0.93]	0.81 [0.71, 0.90]	0.85 [0.79, 0.91]	323.33
	DeiT-Small	0.92 [0.87, 0.96]	0.84 [0.76, 0.91]	**0.85 [0.77, 0.93]**	0.86 [0.80, 0.91]	69.08
CNN	MobileNetV3	**0.94 [0.89, 0.97]**	0.91 [0.84, 0.96]	0.79 [0.69, 0.88]	0.88 [0.82, 0.92]	53.84
	EfficientNetV2	**0.94 [0.90, 0.97]**	0.87 [0.79, 0.93]	0.81 [0.71, 0.90]	0.86 [0.81, 0.91]	**52.89**
	ConvNeXt-Tiny	0.93 [0.88, 0.96]	0.82 [0.74, 0.90]	**0.85 [0.76, 0.93]**	0.85 [0.79, 0.90]	73.10
	DenseNet121	0.92 [0.88, 0.96]	0.91 [0.84, 0.96]	0.76 [0.66, 0.86]	0.86 [0.80, 0.92]	75.26
	InceptionV3	0.91 [0.87, 0.95]	0.87 [0.80, 0.94]	0.81 [0.71, 0.90]	0.86 [0.80, 0.91]	77.49
	ResNet50	0.90 [0.85, 0.94]	0.78 [0.68, 0.86]	0.79 [0.69, 0.89]	0.80 [0.73, 0.86]	58.69
Summary	Median (All)	0.93	0.87 ‡	0.80	0.86	68.57
	Median (CNN)	0.92	0.87 †	0.80	0.86	65.89
	Median (ViT)	0.93	0.88 †	0.80	0.86	68.57
	[c] p-value (Families)	ns	ns	ns	ns	ns

Table 3: The model's sensitivity, specificity, F1-score AUC and single image prediction time are represented in the following table, with their respective confidence intervals (CI). The best model for each element is highlighted in bold. CNN: Convolutional Neural Network, ViT: Vision Transformer.

[a] No statistically significant differences were observed in the AUC values among the ten models (p-value > 0.05).

[b] Cochran's Q Test (Heterogeneity): Used to assess differences in classification performance across models:

‡ Indicates significant global heterogeneity across all 10 models for Sensitivity (p-value < 0.001).

† Indicates significant intra-group heterogeneity within the CNN family (p-value = 0.002) and ViT family (p-value = 0.033) for Sensitivity.

No significant heterogeneity was found for Specificity globally or within subgroups (p-value > 0.05).

[c] Mann-Whitney U Test (Family Comparison): Compared the median performance of the ViT group (n=4) versus the CNN group (n=6). No significant differences were found between the two architectural families for any metric (p-value > 0.05).

Figure 1: Data selection Flowchart, showing exclusion criteria and final distribution following database splitting.

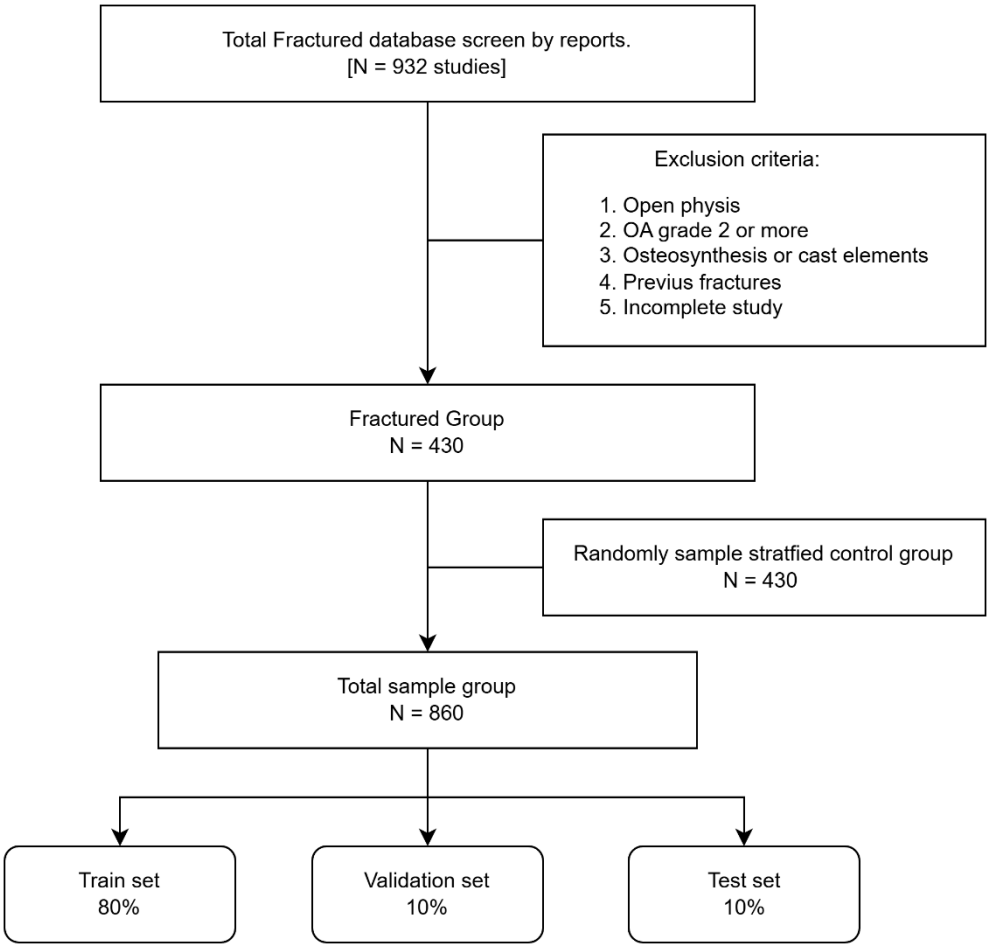


Figure 2: Proposed "Dual-Stream" architecture for knee fracture diagnosis. Each input projection (AP and LAT) is processed by an independent feature extraction (FE) backbone, which can be either a Convolutional Neural Network (CNN) or a Vision Transformer (ViT) model. The resulting feature sequences are concatenated and passed through a Multi-Head Self-Attention (MHA) mechanism to enable cross-view interaction. A Global Average Pooling layer condenses the attention output into a single feature vector, which is then fed into a classification block composed of a Multi-Layer Perceptron (MLP) with dropout regularization. The final output is a fracture probability obtained via a sigmoid activation. Additionally, using the Grad-Cam algorithm, taking the last convolutional layers from the FE blocks, saliency map was obtained for inference and analysis of model interpretation.

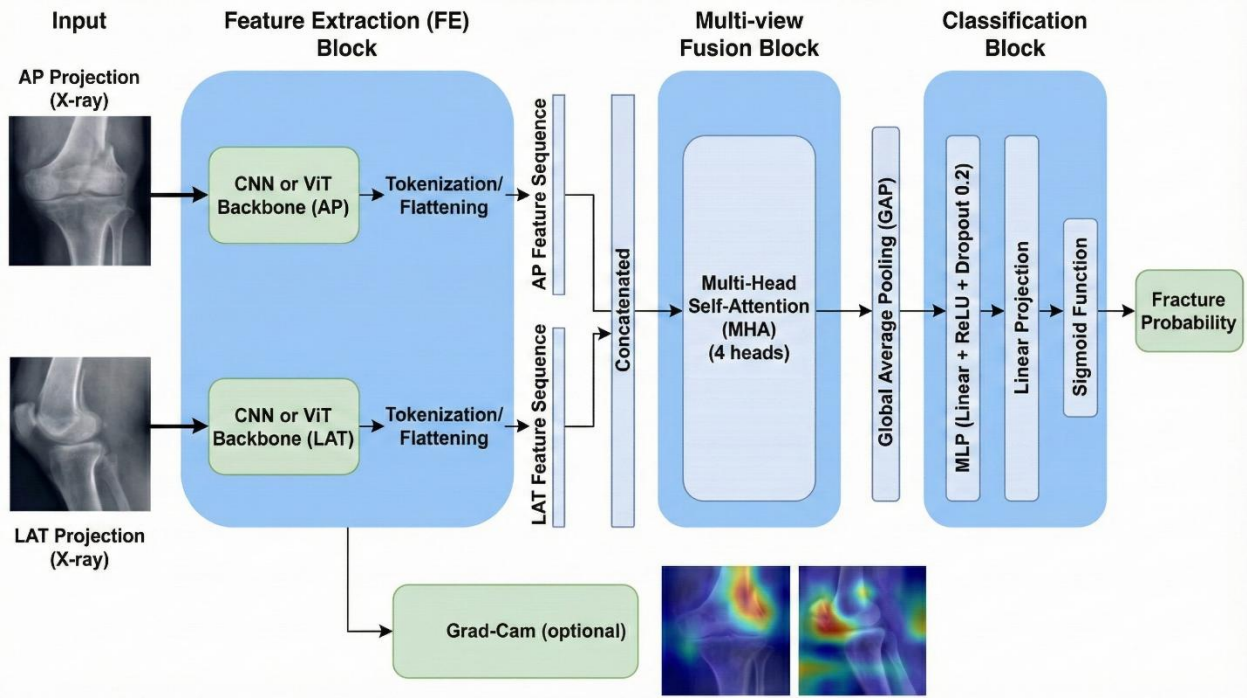


Figure 3: ROC Curves for trained models. No statistically significant differences in AUC were observed among the tested models ($p > 0.05$).

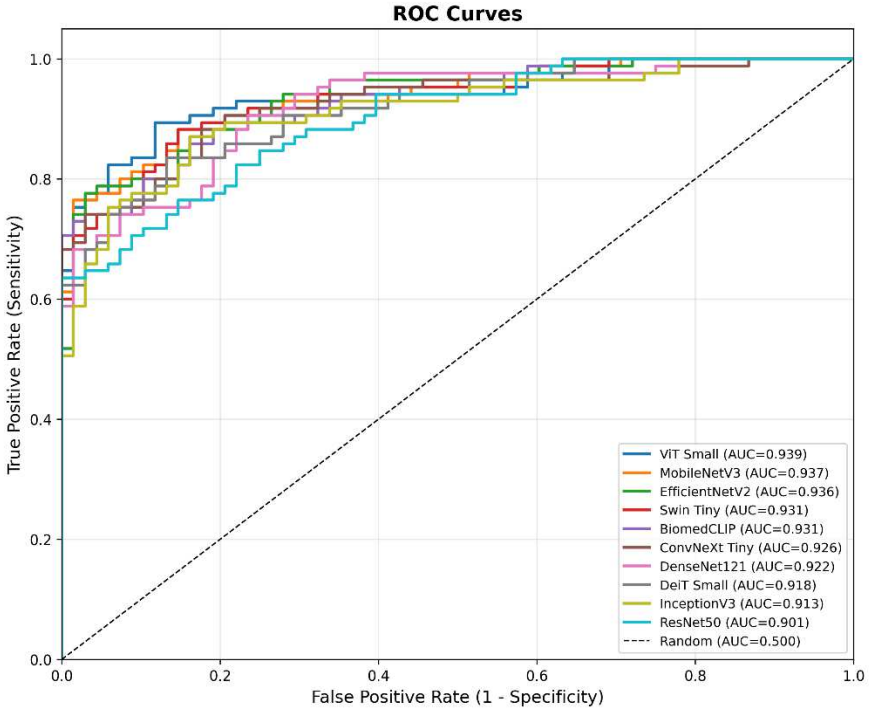
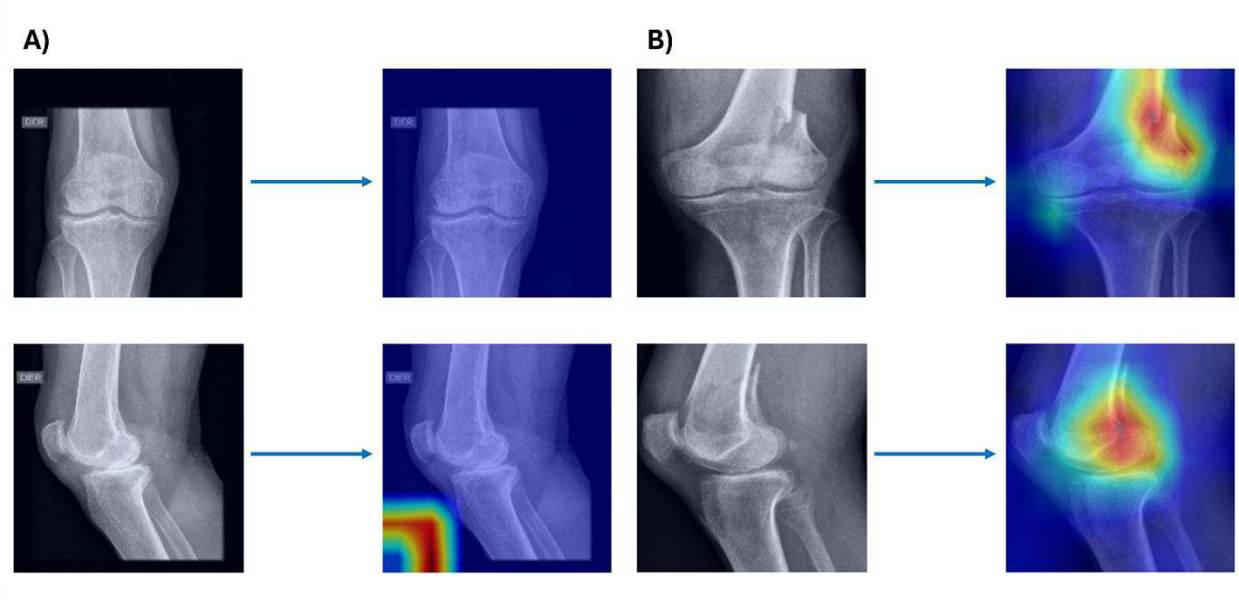


Figure 4: Grad-Cam results example using a ResNet model from two different cases. Top Figure a) shows an example of a True Negative case where the concentration map over the AP image does not highlight any area. Bottom Figure a) shows a lateral view with their respective maps highlighting an area that corresponds to the corner, which probably is interpreted as confusing output. Figure b) shows an example of a True Positive example where the Grad-Cam algorithm correctly identifies the zone of the fracture, identifying a distal femoral fracture in both planes.



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SuplementaryDataChecklistCLAIM.docx