Evaluating Multiple Multiview Fusion Strategies for Knee Fracture Detection in Paired Radiographs

Maximilian Nielsen[1] M.NIELSEN@UKE.DE
Finn Hinnerk Dieckhoff[1,2] FINN.DIECKHOFF@OUTLOOK.COM
Arne Ewald[2] MAIL@AEWALD.NET
Tobias Dust[1] T.DUST@UKE.DE
René Werner[1] R.WERNER@UKE.DE

[1] *Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Martinistr. 52, 20246 Hamburg, Germany*
[2] *NORDAKADEMIE gemeinnützige Aktiengesellschaft Hochschule der Wirtschaft , Köllner Chaussee 11, 25337 Elmshorn, Germany*

Editors: Under Review for MIDL 2026

Abstract

Background: Rapid and accurate evaluation of knee trauma in the Emergency Department is critical. While radiographs are the standard initial assessment, subtle fractures often lack overt visual signs, necessitating computed tomography (CT) for confirmation.

Objective: Unlike prior studies that focus on all types of knee fractures, this work addresses the automatic detection of diagnostically challenging, non-displaced knee fractures and explicitly investigates different view fusion strategies.

Methods: We evaluate multiple multiview deep learning frameworks that leverage complementary information from paired Anterior–Posterior and lateral X-ray projections. To address data scarcity and anatomical complexity, we employ radiology-specific self-supervised pretraining (RAD-DINO) combined with parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA). We systematically evaluate different fusion strategies on a dataset of CT-confirmed knee fracture cases.

Results: Despite the high diagnostic difficulty of the cohort, our best-performing model (Self-Attention Fusion) achieves an AUROC of 0.88.

Conclusion: These findings demonstrate that combining multiview information with domain-adapted pretraining enables robust fracture detection in ambiguous cases.

Keywords: Fracture Detection, Deep Learning, Multiview Fusion, Self-Supervised Learning, Knee Radiography, Emergency Medicine

1. Introduction

Automated interpretation of medical images has increasingly leveraged deep learning to support diagnosis, triage, and clinical decision-making. In the emergency department (ED), rapid and accurate evaluation of musculoskeletal injuries is essential (Bachmann et al., 2024). Patients presenting with suspected knee trauma typically undergo an initial radiographic assessment consisting of paired Anterior–Posterior (AP) and lateral X-ray views. While these radiographs often suffice for clear-cut fractures or confidently normal joints, diagnostically uncertain cases require computed tomography (CT) to confirm or exclude injury (Avci and Kozaci, 2019). To ensure high-fidelity ground truth, we restricted our study cohort to patients with CT-confirmed diagnoses. Consequently, our dataset inherently focuses on ambiguous and difficult cases, reflecting the real-world diagnostic uncertainty of the ED workflow.

This selection bias has crucial implications for algorithm development. A clinically useful AI system is most impactful precisely in these borderline cases, where expert decision-making is challenged by subtle fracture lines, overlapping structures, or poor visibility. Robust multiview models that exploit complementary information from both radiographic projections could therefore serve as valuable decision-support tools, potentially reducing unnecessary radiation exposure, imaging time, and clinical workload.

Research in clinical imaging has predominantly focused on *single-level fusion* strategies, in which each view is processed independently and single view representations are concatenated (Li et al., 2024). However, alternative strategies exist: *early fusion* (combining raw inputs at the pixel level) and *attention based fusion*. The latter enables information flow at deeper network layers by merging latent feature representations via different attention mechanisms found in Transformer architectures (Vaswani et al., 2017). This is particularly relevant for knee trauma, where correlating features across views (e.g., a fat pad sign on lateral view vs. cortical step-off on AP) is critical for diagnosis. In the scope of this study, we compare six different fusion strategies against a single-view benchmark, in which each view is treated as a single independent data point.

To address the challenge of limited labeled data in this specialized domain, we leverage self-supervised pretraining. Recent radiology-focused frameworks, such as RAD-DINO, have demonstrated that learning from large collections of unlabeled radiographs captures domain-specific anatomical structures better than natural-image pretraining (Nielsen et al., 2023). We build upon a RAD-DINO backbone and fine-tune using Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA introduces trainable low-rank matrices into frozen Transformer weights, enabling effective specialization with significantly fewer parameters.

While automated fracture detection has been studied extensively, prior work has largely focused on long-bone injuries (e.g., femoral or humeral fractures) where displacement or cortical discontinuity provide strong visual cues. In contrast, subtle knee fractures in the ED often lack overt radiographic signs. To date, only two studies have focused specifically on knee fractures. Although Lind et al. reported a high AUROC of 0.88, they acknowledged that the inclusion of cases with casts, implants, and other visible cues likely inflated the network's performance (Lind et al., 2021). Furthermore, their ground truth was established via automatic label extraction from text reports rather than definitive CT imaging. Conversely, Van der Gaast et al. utilized only CT-confirmed cases but also included unambiguous fractures, reporting an AUROC of 0.77 (van der Gaast et al., 2025). Crucially, while both studies utilized multiple views, neither explicitly addressed the challenge of multi-view fusion. To our knowledge, this study is the first to concentrate specifically on this diagnostically difficult population and to systematically investigate multi-view fusion strategies for detecting these subtle injuries.

Using our in-house dataset of diagnostically challenging fractures, we demonstrate that combining multiview information with radiology-specific pretraining and LoRA adaptation achieves an AUROC of 0.88. Our findings suggest that these modeling strategies may generalize to other subtle radiographic pathologies where expert disagreement is common (Foroohar et al., 2011; Sayed-Noor et al., 2011).

2. Methods

2.1. Dataset

The study cohort comprises 444 patients presenting to the ED of the University Medical Center Hamburg–Eppendorf (UKE) with suspected acute knee trauma. Each patient underwent a standard radiographic assessment consisting of paired AP and lateral projections, representing the routine diagnostic workflow in acute care. We excluded patients with previous interventions, additional material (e.g., implants), as well as highly arthritic knee joints (Kellgren-Lawrence ¿ 3).

Crucially, study inclusion was restricted to patients who subsequently underwent computed tomography (CT). This requirement ensures that fracture labels are based on high-fidelity CT ground truth rather than X-ray interpretation alone. Consequently, the dataset focuses on cases where clinical ambiguity or injury severity necessitated advanced imaging.

As detailed in Table 1, the dataset was split into training and test sets. The training set consists of 404 patients (229 fractures, 175 non-fractures), while the held-out test set comprises 40 patients (19 fractures, 21 non-fractures). We employed a balanced 90/10 train-test split, with 15% of the training data ($n = 62$) reserved for validation.

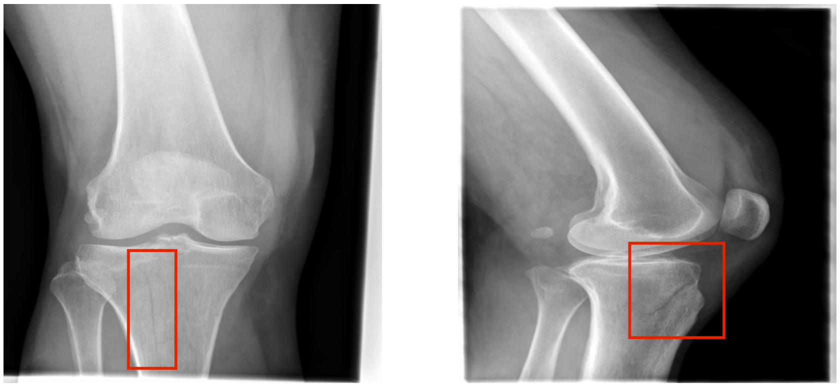
All radiographs underwent a unified preprocessing pipeline to standardize input distribution. Images were resized to 512 × 512 pixels and Z-normalized using global dataset statistics. To enhance local contrast and improve the visibility of trabecular structures, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE). During training, standard data augmentation is performed, namely: cropping, rotation, blurring, inversion, and random deletion of single views. An illustrative image pair is provided in Figure 1.

Table 1: Dataset characteristics

	data set (n=444)
Age (years)	54.8 ± 19.63
Sex	
Female	276
Male	168
Label	
Fracture	250
No Fracture	194

2.1.1. Single View Baseline

To establish a performance baseline, we first train a model that treats each radiographic view as an independent data sample. In this configuration, the network receives a single image as input, regardless of whether it is an AP or lateral view, and predicts a fracture probability $p(y|x)$. During training, paired views from the same patient are decoupled and shuffled, effectively removing any inter-view dependencies. This approach doubles the effective sample size for the baseline model (N=808 training, N=80 test) compared to the



(a) AP

(b) lateral

Figure 1: Exemplary AP (a) and lateral (b) view of a fractured knee (red box indicates fracture site).

multi-view models (N=404 training, N=40 test), allowing us to quantify the performance gain strictly attributable to fusion.

2.2. Multi-View Fusion

We use the single-view baseline to benchmark distinct intermediate fusion strategies, inspired by the taxonomy provided by Li et al. (Li et al., 2024).

Early Fusion In this strategy, integration occurs at the pixel level. The preprocessed AP and lateral radiographs (each $H \times W$) are concatenated along the spatial *width* to form a composite tensor $x \in \mathbb{R}^{H \times 2W}$. This extended input is processed by a single RAD-DINO backbone.

Embedding-Concatenation We employ a simple feature-level fusion strategy. Embeddings extracted from separate views, $z_1, z_2 \in \mathbb{R}^d$, are concatenated along the channel dimension to form a joint representation $z_{\text{fused}} = [z_1 \parallel z_2] \in \mathbb{R}^{2d}$. This vector is passed to the classification head.

Self-Attention To capture non-linear dependencies, we treat the view embeddings as a token sequence and append a learnable classification token, $Z = \{z_{\text{cls}}, z_1, z_2\}$. We apply standard multi-head self-attention (MSA). Classification is performed solely on the output of the context-aware z_{cls} token.

Cross-Attention Unlike self-attention, cross-attention explicitly models the interaction between a primary and supplementary view. We implement this bidirectionally: one branch uses the AP view as the Query (Q) and the lateral as Key (K) and Value (V) to highlight relevant lateral features; the second branch reverses these roles.

Hierarchical-Self Attention This strategy extends self-attention to utilize multi-scale information. In addition to the final output, we extract view-wise ‘[CLS]‘ tokens from intermediate backbone layers (layers 2, 8, and 12). These are aggregated into a single sequence $Z = \{z_{\text{cls}}, z_1^{L2}, z_2^{L2}, \ldots, z_1^{L12}, z_2^{L12}\}$.

Late Fusion We implement a late fusion strategy (Weighted Logits) to isolate decision-level aggregation. Separate classification heads produce logits l_1 and l_2 for each view. These are combined via a learnable parameter α, computing the final output as $l_{\text{final}} = \sigma(\alpha) \cdot l_1 + (1 - \sigma(\alpha)) \cdot l_2$.

2.3. Training & Evaluation

We deploy RAD-DINO, a ViT pretrained with DINOv2 on the MULTI-CXR dataset (Perez-Garcia et al., 2025; Oquab et al., 2023). To mitigate overfitting, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022). Fine-tuning uses the AdamW optimizer ($lr = 5 \times 10^{-4}$, weight decay 0.05, batch size 32). We optimize Cross Entropy loss with label smoothing ($\epsilon = 0.1$). We apply softmax normalization to the output logits to obtain class probabilities. Training is monitored via AUROC on the validation set with early stopping. We report Balanced Accuracy, F1-score, and AUROC on the held-out test set, determining the operating point via Youden's J statistic.

3. Results

Our experimental evaluation focuses on the comparative analysis of six different strategies for multi-view fusion. In general, the results demonstrate the superior predictive capacity of models incorporating multi-view fusion mechanisms compared to the single-view baseline.

Table 2: Overview of experimental results comparing different fusion strategies. The best performance is highlighted in bold. Note that Early Fusion did not converge.

Strategy	**Bal. Acc.**	**F1-Score**	**AUROC**
Single-view (Baseline)	0.77	0.74	0.81
Early Fusion	–	–	–
Embedding-Concatenation	0.75	0.74	0.77
Self-Attention	**0.85**	**0.83**	**0.88**
Cross-Attention	0.78	0.79	0.85
H.-Self-Attention	0.68	0.71	0.81
Late Fusion	0.80	0.79	0.85

Table 2 and Figure 2 summarize the quantitative performance of the evaluated models. Multiple fusion strategies surpassed the baseline across all metrics. The Self-Attention strategy achieved the highest scores in Balanced Accuracy (0.85), F1-Score (0.83), and AUROC (0.88). Other methods, such as Embedding-Concatenation and Hierarchical-Self-Attention, lagged slightly behind.

Among the hierarchical fusion methods, we observed a performance drop as complexity increased. Notably, the Early Fusion approach failed to converge, even after a hyper-

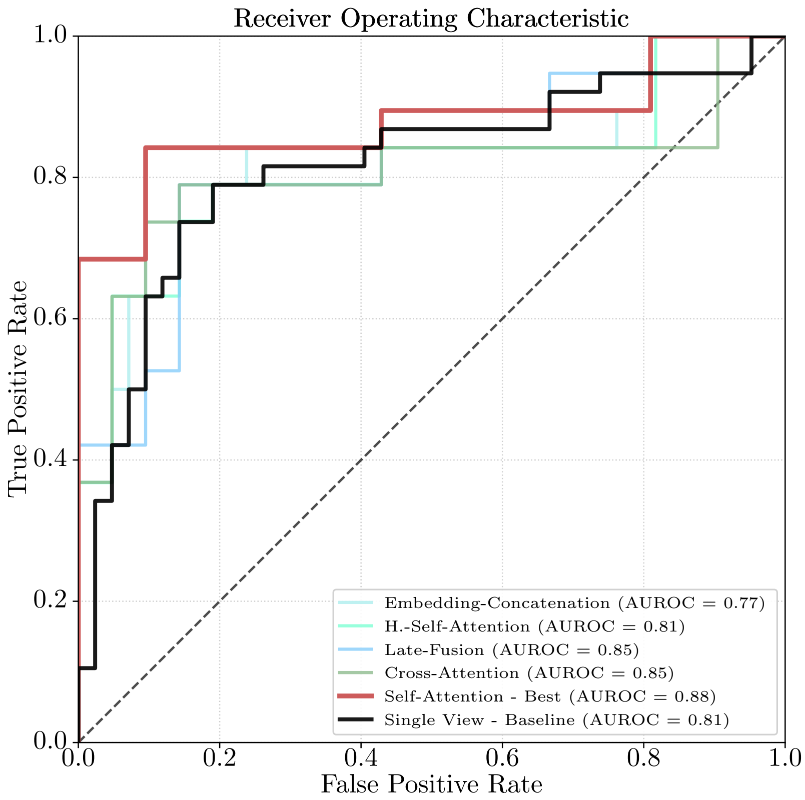
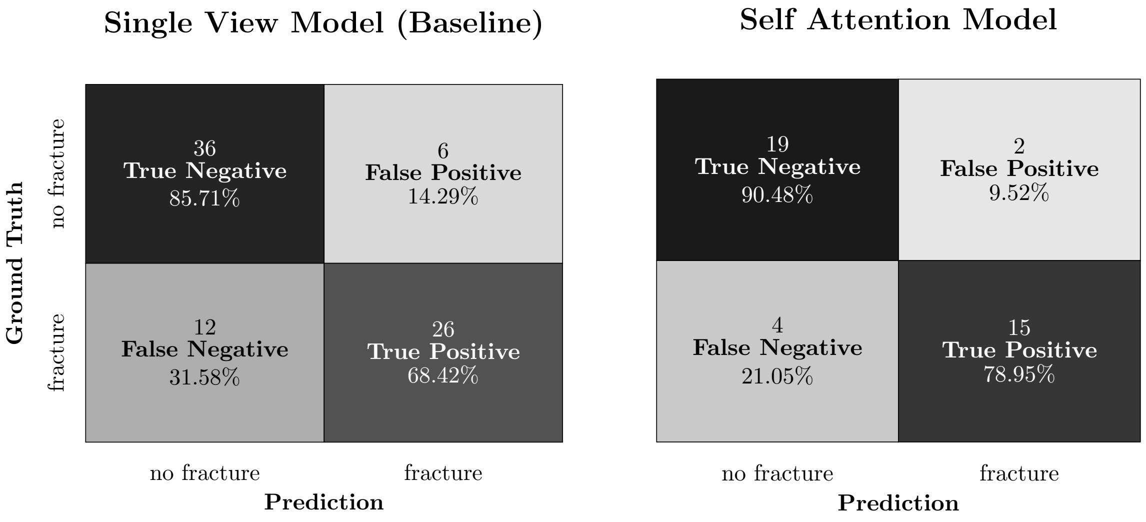


Figure 2: Receiver Operating Characteristic (ROC) curves for all tested configurations. The single-view baseline is depicted in black, and the best-performing model (Self-Attention fusion) in red.

parameter search. We hypothesize that this stems from the input format varying significantly from the original RAD-DINO input format and distribution, rendering the LoRA fine-tuning insufficient to bridge this input shift. Conversely, Late Fusion, despite being the simplest approach, performed robustly, achieving the second-highest AUROC (0.85). Overall, only the Self-Attention-based fusion provided a decisive performance advantage over the other methods. Regarding validation performance, the ranking across all methods is consistent, with validation metrics being, as expected, slightly higher.

Figure 3 compares the confusion matrices for the baseline single-view model (a) and the best-performing self-attention model (b). The difference in absolute sample counts arises from the baseline single-view approach, where image pairs are decoupled, effectively doubling the test-set size. While the baseline model achieved a high true-negative rate (85.71%), it exhibited a substantial false-negative rate (31.58%). In contrast, the self-attention model reduced the false-negative rate to 21.05%, missing only four fractures.

Interestingly the pattern of higher true negative rates is specific for these two models; the remaining models show either more balanced behavior or a clear dominance of true-positive

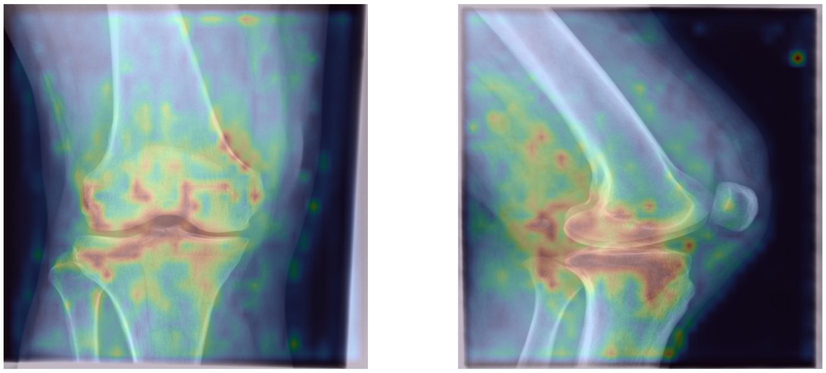


(a)

(b)

Figure 3: Confusion matrices comparing (a) the Single-View Baseline and (b) the best-performing Self-Attention model. Note that sample sizes differ because the Baseline model treats each view as an independent sample (N=80), while the Fusion model processes pairs (N=40).

predictions. This suggests that the performance gains observed with self-attention-based fusion stems primarily from its improved ability to correctly recognize non-fracture cases.



(a) AP

(b) lateral

Figure 4: Exemplary attention visualization of AP (a) and lateral (b) view of fractured knee.

To interpret the decision-making process of the Self-Attention model, we utilized attention rollout visualizations. Figure 4 displays the attention heatmaps overlaid on exemplary

AP and lateral views. The model correctly focuses on the anatomical regions corresponding to the knee joint and fracture site. This localization suggests that the classification is driven by relevant clinical features rather than background artifacts or confounding variables.

4. Discussion

Our study aimed to address the diagnostic challenge of detecting subtle, CT-confirmed knee fractures in an emergency department setting. By systematically benchmarking six fusion strategies, we demonstrated that integrating multi-view radiographic information via Self-Attention mechanisms significantly outperforms the single-view baseline. Our best-performing model achieved an AUROC of 0.88, matching the state-of-the-art reported by Lind et al. (Lind et al., 2021) while operating on a fundamentally more difficult dataset restricted to ambiguous cases.

The superior performance of the Self-Attention strategy (AUROC 0.88, F1 0.83) validates our hypothesis that radiographic views contain complementary feature sets that are best integrated in latent space. By correctly identifying 80̃% of all fractures we demonstrate potential as a safety net for junior clinicians handling difficult cases. Simultaneously, the improved specificity (reduction in False Positives) suggests that such a system could reduce the volume of unnecessary follow-up CT scans, alleviating resource constraints.

Our study is subject to limitations inherent to its design. First, the restriction to CT-confirmed cases, while ensuring high-fidelity ground truth, introduces a selection bias that excludes unambiguous fractures. Consequently, while our model demonstrates robustness on diagnostically difficult cases, its performance on overt fractures remains formally untested in this cohort. Second, the relatively small size of our test set limits the statistical power of our evaluation and warrants validation on larger, multi-center datasets to ensure generalizability. Future work should expand beyond algorithmic metrics to clinical relevance, specifically through a reader study comparing the model against clinical personnel of varying expertise (e.g., emergency physicians vs. board-certified radiologists). Furthermore, prospective evaluation is needed to determine if interactive attention maps can effectively guide clinician focus and reduce diagnostic errors in real-time workflows.

5. Conclusion

We present the first targeted investigation into multi-view deep learning fusion strategies for detecting subtle, diagnostically challenging knee fractures. We show that a Self-Attention fusion architecture, fine-tuned on a foundation model backbone, effectively synthesizes complementary radiographic information. This approach sets a new benchmark for ambiguous knee trauma cases, offering a tangible path toward reducing missed diagnoses in the emergency department.

References

Mustafa Avci and Nalan Kozaci. Comparison of x-ray imaging and computed tomography scan in the evaluation of knee trauma. *Medicina*, 55(10):623, 2019.

Rikke Bachmann, Gozde Gunes, Stine Hangaard, Andreas Nexmann, Pavel Lisouski, Mikael Boesen, Michael Lundemann, and Scott G Baginski. Improving traumatic fracture detection on radiographs with artificial intelligence support: a multi-reader study. *BJR—Open*, 6(1):tzae011, 2024.

Abtin Foroohar, Rick Tosti, John M Richmond, John P Gaughan, and Asif M Ilyas. Classification and treatment of proximal humerus fractures: inter-observer reliability and agreement across imaging modalities and experience. *Journal of orthopaedic surgery and research*, 6(1):38, 2011.

Edward J Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. *ArXiv preprint arXiv:2106.09685*, 2022.

Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, 2024.

Anna Lind, Ehsan Akbarian, Simon Olsson, Hans Nåsell, Olof Sköldenberg, Ali Sharif Razavian, and Max Gordon. Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 ao/ota classification system. *PLoS One*, 16(4): e0248809, 2021.

Maximilian Nielsen, Laura Wenderoth, Thilo Sentker, and René Werner. Self-supervision for medical image classification: State-of-the-art performance with˜ 100 labeled training samples per class. *Bioengineering*, 10(8):895, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Fernando Perez-Garcia, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

Arkan S Sayed-Noor, Per-Henrik Ågren, and Per Wretenberg. Interobserver reliability and intraobserver reproducibility of three radiological classification systems for intra-articular calcaneal fractures. *Foot & ankle international*, 32(9):861–866, 2011.

N van der Gaast, P Bagave, N Assink, S Broos, RL Jaarsma, MJR Edwards, E Hermans, FFA IJpma, AY Ding, JN Doornberg, et al. Deep learning for tibial plateau fracture detection and classification. *The Knee*, 54:81–89, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.