Arthritis & Rheumatology

AN OFFICIAL JOURNAL OF
THE AMERICAN COLLEGE OF
RHEUMATOLOGY

Expanding from Unilateral to Bilateral: a robust deep learning-based approach for radiographic osteoarthritis progression

Journal:	*Arthritis & Rheumatology*
Manuscript ID	Draft
Wiley - Manuscript type:	Full Length
Date Submitted by the Author:	n/a
Complete List of Authors:	Yin, Rui; Nanjing Medical University affiliated Nanjing Hospital, Department of Sports Medicine and Joint Surgery Chen, Hao; University of Birmingham, School of Computer Science Tao, Tianqi; Nanjing Medical University affiliated Nanjing Hospital, Department of Sports Medicine and Joint Surgery Zhang, Kaibin; Nanjing Medical University affiliated Nanjing Hospital, Department of Sports Medicine and Joint Surgery Yang, Guangxu; The Forth Affiliated Hospital of Nanjing Medical University, Department of Orthopedic Surgery Shi, Fajian; The Fourth Affiliated Hospital of Nanjing Medical University, Department of Orthopedic Surgery Jiang, Yiqiu; Nanjing Medical University affiliated Nanjing Hospital, Department of Sports Medicine and Joint Surgery Gui, Jianchao; Nanjing Medical University affiliated Nanjing Hospital, Department of Sports Medicine and Joint Surgery
Keywords:	Osteoarthritis, Deep Learning, X-ray, Risk Assessment, Cross Attention
<B>Disease Category</b>: Please select the category from the list below that best describes the content of your manuscript.:	Osteoarthritis

Expanding from Unilateral to Bilateral： a robust deep learning-based approach for radiographic osteoarthritis progression

Rui Yin, PhD[1], Hao Chen, MS[2], Tianqi Tao, MS[1], Kaibin Zhang, MS[1], Guangxu Yang, MD[3], Fajian Shi, MD[3], Yiqiu Jiang, MD[1], Jianchao Gui, MD[1*]

Affiliations:

[1] Department of Sports Medicine and Joint Surgery, Nanjing Medical University affiliated Nanjing Hospital, Nanjing, China

[2] School of Computer Science, University of Birmingham, Birmingham, UK

[3] Department of Orthopedic Surgery, The Fourth Affiliated Hospital of Nanjing Medical University, Nanjing, China

Corresponding information for Corresponding Author:

*Jianchao Gui

Professor and chief of the department

Department of Sports Medicine and Joint Surgery, Nanjing Medical University affiliated Nanjing Hospital Address: Changle Road 68, Nanjing 210006, China

Telephone number: 025-52271250

E-mail: gui1997@126.com

Running title: BikNet for OA progression prediction

Funding: This study was supported by the National Natural Science Foundation of China (no. 81672210).

Conflict of interest: The authors declare no competing interests.

Abstract

Objective: To develop and validate a deep learning model for predicting osteoarthritis (OA) progression based on bilateral knee joint views.

Methods: In this retrospective study, knee joints from bilateral posteroanterior knee radiographs of participants in the Osteoarthritis Initiative were analyzed. At baseline, participants were divided into testing set 1 and development set according to the different enrolled sites. The development set was further divided into a training set and a validation set in an 8:2 ratio for model development. At 48-month follow-up, eligible patients were formed testing set 2. The Bilateral Knee Neural Network (BikNet) was developed using bilateral views, with the knee to be predicted as the main view and the contralateral knee as the auxiliary view. DenseNet and ResNext were also trained and compared as the unilateral model. Two reader tests were conducted to evaluate the model's value in identifying "early-stage" OA.

Results: Totally 3583 participants were evaluated. The BikNet we proposed outperformed ResNext and DenseNet (all AUC < 0.71, P < 0.001) with AUC values of 0.761 and 0.745 in testing sets 1 and 2, respectively. With assistance of the BikNet model increased clinicians' sensitivity (from 28.1-63.2% to 42.1-68.4%) and specificity (from 57.4-83.4% to 64.1-87.5%) of "early-stage" OA diagnosis and improved inter-observer reliability.

Conclusion: The deep learning model, constructed based on bilateral knee views, holds promise for enhancing the assessment of OA and demonstrating greater robustness during subsequent follow-up evaluations. BikNet represents a potential tool or imaging biomarker for predicting OA progression.

Introduction

Osteoarthritis (OA) is the leading cause of chronic disability in the United States and one of the fastest-growing medical conditions worldwide (1,2). With aging populations, the incidence of OA is expected to rise even further in the coming years. Despite its considerable impact on public health, no disease-modifying drug therapy for OA has received approval from regulatory agencies such as the US Food and Drug Administration or the European Medicines Agency (3,4). The uncertain progression of OA poses a significant challenge in designing clinical trials, as only a tiny proportion of patients (4-8%) are likely to experience radiographic progression within four years (5). Including patients predisposed to progression or in the early stages of the disease in clinical trial cohorts can accelerate drug development for OA and advance personalized and precision-targeted interventions (6).

X-ray is a commonly used and cost-effective method for assessing OA due to its convenience. However, hand-crafted radiographic features have limited value in facilitating early diagnosis and predicting disease progression (7,8). Recently, deep learning (DL) has emerged as a promising technique for medical image analysis. DL models heuristically learn important features from images to enable accurate clinical predictions, circumventing the need for laborious manual feature engineering and surpassing the performance of conventional machine learning methods (9–11). Prior studies demonstrated the feasibility of using DL analysis of baseline radiographs to predict knee pain (12,13), medial joint space loss (14), and subsequent total knee arthroplasty (TKA) in OA patients (15). Although DL has shown impressive performance in predicting OA-related outcomes, previous works primarily focused on analyzing each knee individually, overlooking the systemic nature of the disease and the potential influence of the contralateral joint. Given the high prevalence of bilateral knee OA, clinicians need to account for both knees concurrently when assessing the relationship between symptoms, physical function, and structural disease, as should DL models do (16–18). Moreover, since OA is a chronic condition that necessitates ongoing follow-up and reassessment (1,2), it is critical to evaluate the models' performance in follow-up scenarios.

In order to overcome the limitations of previously reported DL models for OA, we proposed the Bilateral Knee Neural Network (BikNet), which incorporates cross-attention (19–21). The cross-attention mechanism in the BikNet enables the network to evaluate both knees simultaneously and learn their interdependence. This capability allows it to capture and leverage information from bilateral views from the raw X-ray, resulting in more precise predictions. Our hypothesis was that BikNet could learn more effective representations and achieve superior performance compared to previous DL models (unilateral models) that evaluate only one knee at a time in predicting the progression of OA at both baseline and subsequent follow-up time points. Furthermore, we contend that BikNet can aid in diagnosing "early-stage" OA or predicting OA onset.

Methods

Datasets. This retrospective study analyzed 12,650 knees using 6,325 radiographs obtained from 3,585 participants of the Osteoarthritis Initiative (OAI), a multicenter prospective study (https://nda.nih.gov/oai/). All individuals were recruited consecutively from February 2004 to May 2006. A total of 1,211 participants were excluded for various reasons, such as knee replacement, rheumatoid arthritis, at least one knee with Kellgren-Lawrence grade (KLG) 4, or missing follow-up for 48 months (as shown in Figure S1). Baseline radiographs (n=3,585) were utilized for both model development and testing. The participants were initially divided into a development set (from B, C, or D) and a testing set 1 (from A or E) based on the enrolled site. The development set was then randomly split into training and validation sets of 80% (n=2,227) and 20% (n=557), respectively. To further evaluate the models' robustness and mimic clinical scenarios, testing set 2 (n=2,653) was created by obtaining 4-year follow-up radiographs. Participants were recruited at four clinical sites, and the Health Insurance Portability and Accountability Act–complaint study was approved by the institutional review board (IRB) at each site. All participants gave written informed consent.

In this study, nonprogression was defined as no change in KLG or a change from KLG 0 to KLG 1, while progression was defined as an increase in KLG of at least one or the receipt of TKA during the follow-up period (22).

Deep learning workflow. The deep learning workflow is depicted in Figure 1. In brief, images of all participants were cropped and preprocessed to fit the model inputs. The BikNet was trained using a multitask paradigm with two auxiliary tasks. Subsequently, the model's output and heatmap could be utilized to aid clinical OA evaluation. All deep learning models were trained on a workstation equipped with an Nvidia Tesla A100 and an Intel Xeon Gold 5215 CPU. Further details are summarized below.

Image preprocessing. Before feeding the images into the model, several preprocessing steps were performed sequentially to normalize the dataset, as demonstrated in Figure S2. First, a pre-trained Hourglass network (23) was utilized to extract a 700 × 700 pixel region of interest (ROI) from all bilateral posteroanterior fixed-flexion knee radiographs at baseline and the 48-month follow-up time points. The left knee images were flipped to the right knee configuration and resized to 310 × 310 pixels after undergoing quality control by radiologists. Next, the images underwent histogram clipping between the 5th and 99th percentiles, followed by global contrast normalization, wherein the minimum image value was subtracted from all image pixels, and the resulting values were divided by the maximum pixel value. Lastly, histogram normalization was carried out to improve the recognition accuracy by enhancing the characteristics of the trabecular bone texture (24).

Model architecture. The diagram of our model's architecture is illustrated in Figure 2. In contrast to previous studies that take each knee as an isolated input, we take inspiration from how clinicians naturally diagnose patients and present our BikNet, which can leverage information gained from bilateral views. In our model, the knee to be evaluated serves as the main view, while the contralateral knee serves as an auxiliary view to provide complementary information to improve prediction accuracy. To better fuse cross-view features, we designed a cross-attention module to serve as an inquiry mechanism. This module generates a query vector for each view to indicate which part of the feature from the counterpart is more important to the prediction.

Our study employed a multitask learning paradigm to predict both OA progression, as well as the auxiliary tasks of OA diagnosis and anatomical landmarks identification. The auxiliary tasks could serve as a regularization measure to help the model focus on the key structure, particularly features from the contralateral view, and improve performance, robustness and training speed of the network (25). The OA diagnosis task involved classifying cases as either OA or non-OA based on the current KLG, where a KLG ≥ 2 was defined as OA. Meanwhile, the task of anatomical landmarks identification was a regression task aimed at predicting seven key landmarks in the tibiofibular joint. These landmarks included the midpoint of the intercondylar notch of the femur, the intercondylar eminence of the tibia, and the edges of the joint. As the primary focus of our study was on the main task of predicting OA progression, we did not include a detailed discussion of the results of the auxiliary tasks, which were added solely to improve network optimization during training.

More details and the bilateral hypothesis justification can be found in Appendix E1. The code and model are available at https://github.com/chqwer2/Bilateral-Knee-Network

Model Comparison and Visualization. To demonstrate the superiority of the model developed on bilateral knee views, we compared it with the best-performing backbones (DenseNet and ResNext) from previous studies that predicted OA progression, which served as unilateral convolutional neural network (CNN) models (14,26). The result reported by Panfilov et al. (27) was adopted as a benchmark since it had been the previous state-of-the-art method and used the same definition of OA progression as we did. Additionally, we evaluated commonly used DL models in medical imaging, including ResNet34, ResNet50, and EfficientNet, to supplement our analysis (12,15,24,28). All models were trained on the training set and evaluated on two separate testing sets to assess their predictive performance at the patients' baseline and follow-up visits. Evaluation metrics, including the area under the curve (AUC), sensitivity, and specificity, were used to assess the models' performance.

To provide a human-readable interpretation of the DL model, we utilized a class activation map (CAM) technique to identify the regions where the model focused its attention and discern how it learned discriminative features for risk scores (29,30).

Reader Test. Differentiating individuals with an impending onset of disease, referred to as "early-stage" OA, is crucial for identifying patients who require preventive care and has real potential to better define OA subgroups (6,31). In this study, we defined knees belonging to the early-stage OA group as those without radiographic OA (KLG 0-1) at baseline but showed progression of one or more KLGs (KLG $\geq$ 2) over a four-year period. We conducted two experiments to evaluate the performance of our model in assisting with the diagnosis of "early-stage" OA. In Experiment 1, seven experienced clinicians, including four orthopedists and three radiologists, were given only bilateral posteroanterior fixed-flexion knee radiographs and asked to predict if a patient would experience the onset of OA within 48 months and which knee would be affected. In Experiment 2, clinicians were provided with heat maps and model output, in addition to plain radiographs, to improve their predictions. For both experiments, we randomly selected 200 raw radiographs from 200 participants, of which 50 were "early-stage" OA cases (57 among 400 knees), from two testing sets. All reading experiments were performed on diagnostic computer monitors. Figure S3 displays the interface utilized by clinicians to evaluate the risk of OA onset.

Statistical Analysis. Statistical analysis was performed using R (version 4.02). All analyzed data consisted of statistically independent observations. A P-value less than 0.05 was considered statistically significant. To assess the predictive performance of BikNet and unilateral CNN models in two hold-out testing datasets, receiver operator characteristic (ROC) analysis was used, and the AUCs were calculated. Standard deviations and 95% confidence intervals (CI) were obtained through bootstrapping with 2,000 redraws unless otherwise stated. The Youden index was used to determine optimal model sensitivity and specificity. The DeLong test (32) was used to compare the AUCs of the BikNet and unilateral models. Inter-observer agreement between the seven clinicians was evaluated in the reader test using Fleiss' κ.

Results

Subject Characteristics. The participants had a mean age of 60.8 ± 9.17 years and a mean body mass index (BMI) of 28.3 ± 4.79 kg/m² at baseline. Among the 3,583 participants, 2,161 were women, which accounted for 59.3% of the sample. In the subsequent follow-up period (testing set 2), the mean age of the 2,653 participants was 64.2 ± 9.00 years, with 1,573 of them (59.3%) being women. The percentages of progression of OA were 13.9%, 11.0%, 14.0%, and 7.2% in the training, validation, testing 1, and testing 2 datasets, respectively. Table 1 provides an overview of the participant characteristics and summarizes the grades and frequencies of radiographic OA features.

Model assessment and comparison for OA progression prediction. Table 2 presents the results of using Panfilov et al. (27) as the benchmark for our study. They achieved an AUC of 0.71 using ResNext as the backbone. Despite slight differences in participant selection and image preprocessing, the performance of the ResNext unilateral model reported in our study is comparable to theirs (AUC: 0.707 vs. 0.71), supporting our adoption of their outcomes as a reference and the fairness of comparing BikNet with unilateral models. The ROC curve analysis of BikNet is presented in Figure 3A-B. In testing set 1, BikNet exhibited superior performance with an AUC of 0.761 [0.728-0.795], outperforming ResNext (0.707 [0.670-0.743], P < 0.001), DenseNet (0.708 [0.669-0.744], P < 0.001), and the benchmark (0.71). BikNet also achieved the highest AUC in testing set 2 with a value of 0.746, compared to ResNext (0.667 [0.640-0.693], P < 0.001) and DenseNet (0.649 [0.621-0.677], P < 0.001). In testing set 1, the sensitivity and specificity of BikNet were 0.665/0.774, compared to 0.746/0.556 and 0.518/0.805 for ResNext and DenseNet, respectively. In testing set 2, the sensitivity and specificity of BikNet, ResNext, and DenseNet were 0.675/0.738, 0.788/0.481, and 0.702/0.521, respectively. Unlike unilateral models, BikNet achieved a balance between sensitivity and specificity (Table 2). BikNet significantly outperformed other commonly used backbones as well, including ResNet34 (AUC: 0.681/0.651, all P < 0.001), ResNet50 (AUC: 0.699/0.646, all P < 0.001), and EfficientNet (AUC: 0.655/0.652, all P < 0.001). Detailed results of the comparison with other backbone models can be found in Figure S4 and Table S1.

Assistance in the diagnosis of early-stage OA. To assess the effectiveness of our model in assisting clinicians with the detection of "early-stage" OA, we conducted two reader tests. In the first experiment, most clinicians were unable to reliably differentiate between the two groups, except for one orthopedic surgeon (F.J) with over 20 years of experience in joint surgery. It was found that the performance among clinicians varied significantly, with sensitivity ranging from 28.1% to 63.2% and specificity ranging from 57.4% to 83.4% (Table S2). This was expected as the current approach did not enable clinicians to diagnose "early-stage" OA. In the second test, results improved substantially with the additional informative presentation of the model predictions. As shown in Table S2, both sensitivity and specificity consistently improved, ranging from 42.1% to 68.4% and 64.1% to 87.5%, respectively. Furthermore, all clinicians achieved much better performance, as quantified by the ROC-AUC (Figure 3C-D). It was also noteworthy that AI support helps clinicians rate radiographs more consistently. Fleiss' kappa was 0.203 for Experiment 1, while the agreement between clinicians was higher in Experiment 2, with a kappa of 0.365 (see Table S3).

Interpretation and visualization for the BikNet. Gradient-weighted CAM after the last convolutional layer of the model was overlaid with the radiograph to show the relevance of specific areas for the model classification. The results are presented in Figure 4, which indicates that the model mainly focused on regions near the joint space to learn features related to the knee and classify samples between the two groups. For progression OA (Figure 4A), the model's attention was primarily on the medial joint space or osteophytes, while for nonprogression OA (Figure 4B), the attention was distributed over the joint space with low specificity. These findings suggest that the model learned to assess relevant features rather than just image correlations. Figure 4C illustrates examples of prediction errors caused by poor image quality and obscured bony structures.

Discussion

Our study presents a fully automated deep learning-based system for predicting the progression of OA by evaluating bilateral joint views concurrently on radiographs. Specifically, the system uses the knee under evaluation as the main view and the contralateral joint as the auxiliary view to mimic the evaluation approach used by clinicians. The proposed DL model, named BikNet, achieved outstanding results with AUCs above 0.745 in both baseline (testing set 1) and follow-up (testing set 2) stages. Moreover, BikNet considerably enhanced the sensitivity and specificity of "early OA diagnosis" by clinicians, highlighting the promising potential of computer-based methods for evaluating OA.

Although radiographic features have limited added value in predicting the progression of OA, previous studies have confirmed the potential of DL in assessing OA using radiographs. Guan et al. (14) utilized a DenseNet model to predict medial joint space loss and reported higher performance of DL models based on knee X-rays compared to traditional models using demographic and radiographic risk factors. Tiulpin et al. (26) proposed an OA prediction model based on ResNext, achieving a 6% higher accuracy in identifying progressive cases during a 60-month follow-up period than previously used methods in OA associated literature. Panfilov et al. (27) extended Tiulpin's approach and reported an AUC of 0.71 for a DL method based on X-ray in predicting OA progression, using the same definition of progression as in our study. However, prior studies on DL for OA have focused on each joint as a single entity, whereas knee OA typically affects both joints in the absence of local risk factors. Metcalfe et al. (18) reported that almost 80% of patients with unilateral disease at baseline developed bilateral OA during a 12-year follow-up, while Cotofana et al. (17) found that the risk of OA in "normal knees" is strongly related to the contralateral joint OA status. Therefore, it is crucial to explore a more reasonable DL architecture that can assess bilateral knees simultaneously, which is precisely the objective of BikNet.

Our model simultaneously takes both views as input and fuses them using a cross-attention module. A query feature vector is generated for each view in this module, which is sent to another view for establishing mapping from one view to another. The query vector is then answered by another view through the establishment of a cross-view directed relationship. To improve the model's performance, we designed our model to simulate a clinical diagnosis process, where a doctor first identifies anatomical landmarks, assesses joint space narrowing, and measures the knee alignment to predict the potential risk of OA progression. Our method utilized multitask learning to design auxiliary tasks that explicitly predict the OA diagnosis and landmarks to mimic the abovementioned progress. Although direct comparison of our model with most prior studies is challenging due to differences in outcome definition and cohort selection, BikNet outperformed previously reported backbone models in the same setting and surpassed the benchmark reported by Panfilov et al. (27) at the baseline time point. Moreover, as a chronic disease, ongoing follow-up is needed for OA (1,2). As we know, we were the first to externally validate the OA-related models' performance in the follow-up scenario. It was not surprising that the performance of unilateral CNN models declined significantly and showed weak discrimination during follow-up. However, due to the effective fusion of the features from the contralateral view, BikNet maintained a fair discrimination ability, demonstrating superior robustness compared to models based on unilateral views.

Recent studies have shown the potential benefits of DL-aided systems for various clinical applications. For instance, McKinney et al. (33) developed a DL model for diagnosing breast cancer and reported that their model outperformed six radiologists. Similarly, Kim and colleagues conducted a reader study to assess the performance of radiologists when examining mammograms with or without the assistance of a DL algorithm (34). Their results showed that the diagnostic accuracy of radiologists was significantly enhanced when working in conjunction with DL. In one recent review, Foster et al. (6) noted that informatics systems and clinical decision tools are starting to incorporate OA-related predictive models to facilitate shared decision-making. We conducted two reader experiments to evaluate the assistance of BikNet in "early-stage" OA diagnosis. It was found that neither radiologists nor orthopedists were able to identify patients who were susceptible to developing OA when given only raw X-rays and clinical information (Figure S3A). However, when presented with additional informative visuals, such as heatmaps and model prediction, the performance of all clinicians improved substantially. Specifically, both sensitivity and specificity consistently improved to ranges of 42.1-68.4% and 64.1-87.5%, respectively, and all clinicians achieved better performance as quantified by the ROC curve (see Figure 3C-D). Given that prognosticating OA remains challenging despite extensive clinical and scientific research efforts, identifying patients who are in the early-stage of OA or experiencing OA progression is of paramount importance to guide treatment and potentially facilitate new preventive or curative treatment strategies. With the assistance of our DL approach, clinicians may have the potential to identify patients with "early-stage" OA based only on clinical information and X-rays.

While our initial results are promising, further technical development and validation are necessary before our DL model can be implemented in clinical practice. The radiographic data included in the OAI were obtained using standardized methods across sites and regularly reviewed for quality by the OAI Quality Assurance Center. However, there is still variation in image quality that can affect the training of DL models (15). This variation would make it more challenging to train the DL model accurately and generalize its performance to test datasets. Additionally, DL model performance declined over time, as mentioned above, when evaluating subsequent follow-up data due to disease progression and image quality changes, particularly for unilateral models. These factors can ultimately affect the reliability and validity of the model in real clinical practice. Therefore, future studies should focus on developing more robust and generalizable DL models that can handle variations in image quality and disease progression over time (35,36). Additionally, the current BikNet has been designed specifically for X-ray imaging considering the cost-effectiveness and convenience in clinical practice. However, it has been demonstrated that magnetic resonance imaging (MRI) based DL model or integrating MRI and X-ray can further enhance the performance of OA progression prediction (increasing AUC from 0.71 to 0.76) (27,37,38). In spite of this, BikNet achieved comparable performance with multimodal models by efficiently learning and integrating features from the contralateral joint. We plan to explore the feasibility and effectiveness of a multimodal BikNet in further work. Moreover, it is important to note that BikNet should not be considered an autonomous diagnostic approach, but rather an imaging biomarker or risk assessment tool. It should be utilized in conjunction with other factors, such as clinical risk factors, biochemical markers, multi-omics data, or other modality images, to aid in the assessment of OA, as demonstrated in the reader test.

Our study has several limitations. Firstly, the data utilized was obtained solely from the OAI, which has a limited representation of the Asian population (15). Therefore, it is necessary to validate the efficacy of BikNet further using data from different racial groups. Furthermore, the progression was defined as an increase in KLG within 48 months, which is the most widely accepted definition (22). However, the difference in definition means that our model cannot be directly compared with some previous models (14,26). Nevertheless, we attempted to make a fair comparison by incorporating the best-performing backbone networks used previously in constructing the unilateral model. Our reproduced unilateral model achieved performance similar to that reported by Panfilov et al. (AUC: 0.707 vs. 0.71), indirectly validating the efficacy of this comparative approach.

Conclusion

In conclusion, the current study demonstrated the practicability and efficacy of utilizing bilateral knee views for predicting OA progression. The proposed BikNet outperformed previous unilateral models and enabled us to construct an effective DL model by incorporating features from the contralateral joint. Our model mimics the way clinicians evaluate patients and enhances the reliability. Additional validation during follow-up time points and reader tests further emphasized the robustness of BikNet in clinical scenarios. Moreover, this approach may have the potential for generalization to the assessment of other systemic diseases that involve bilateral limbs, such as rheumatoid arthritis.

Reference

1. Sharma L. Osteoarthritis of the Knee. Solomon CG, ed. *N Engl J Med* 2021;384:51–59.

2. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. *The Lancet* 2019;393:1745–1759.

3. Yazici Y, McAlindon TE, Gibofsky A, Lane NE, Clauw DJ, Jones MH, et al. Results from a 52-week randomized, double-blind, placebo-controlled, phase 2 study of a novel, intra-articular wnt pathway inhibitor (SM04690) for the treatment of knee osteoarthritis. *Osteoarthritis and Cartilage* 2018;26:S293–S294.

4. Eckstein F, Hochberg MC, Guehring H, Moreau F, Ona V, Bihlet AR, et al. Long-term structural and symptomatic effects of intra-articular sprifermin in patients with knee osteoarthritis: 5-year results from the FORWARD study. *Annals of the Rheumatic Diseases* 2021;80:1062–1069.

5. Driban JB, Harkey MS, Barbe MF, Ward RJ, MacKay JW, Davis JE, et al. Risk factors and the natural history of accelerated knee osteoarthritis: a narrative review. *BMC Musculoskelet Disord* 2020;21:332.

6. Foster NE, Eriksson L, Deveza L, Hall M. Osteoarthritis year in review 2022: epidemiology & therapy. *Osteoarthritis and Cartilage* 2023:S1063458423007306.

7. Runhaar J, Kloppenburg M, Boers M, Bijlsma JWJ, Bierma-Zeinstra SMA. Towards developing diagnostic criteria for early knee osteoarthritis: data from the CHECK study. *Rheumatology (Oxford)* 2020;60:2448–2455.

8. Wang Q, Runhaar J, Kloppenburg M, Boers M, Bijlsma JWJ, Bierma-Zeinstra SMA. Diagnosis for early stage knee osteoarthritis: probability stratification, internal and external validation; data from the CHECK and OAI cohorts. *Seminars in Arthritis and Rheumatism* 2022;55:152007.

9. Chen X, Wang X, Zhang K, Fung K-M, Thai TC, Moore K, et al. Recent advances and clinical applications of deep learning in medical image analysis. *Med Image Anal* 2022;79:102444.

10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis* 2017;42:60–88.

11. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med* 2022;28:1773–1784.

12. Guan B, Liu F, Mizaian AH, Demehri S, Samsonov A, Guermazi A, et al. Deep learning approach to predict pain progression in knee osteoarthritis. *Skeletal Radiol* 2022;51:363–373.

13. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med* 2021;27:136–140.

14. Guan B, Liu F, Haj-Mirzaian A, Demehri S, Samsonov A, Neogi T, et al. Deep learning risk assessment models for predicting progression of radiographic medial joint space loss over a 48-MONTH follow-up period. *Osteoarthritis and Cartilage* 2020;28:428–437.

15. Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, et al. Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative. *Radiology* 2020;296:584–593.

16. Messier SP, Beavers DP, Herman C, Hunter DJ, DeVita P. Are unilateral and bilateral knee osteoarthritis patients unique subsets of knee osteoarthritis? A biomechanical perspective. *Osteoarthritis and Cartilage* 2016;24:807–813.

17. Cotofana S, Wirth W, Kwoh KC, Hunter DJ, Duryea J, Eckstein F. Is the risk of incident radiographic knee OA related to severity of contra-lateral radiographic knee status? -data from the osteoarthritis initiative. *Osteoarthritis and Cartilage* 2013;21:S58–S59.

18. Metcalfe AJ, Andersson ML, Goodfellow R, Thorstensson CA. Is knee osteoarthritis a symmetrical disease? Analysis of a 12 year prospective cohort study. *BMC Musculoskelet Disord* 2012;13:153.

19. Chen C-F (Richard), Fan Q, Panda R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In: ; 2021:357–366. Available at: https://openaccess.thecvf.com/content/ICCV2021/html/Chen_CrossViT_Cross-Attention_Multi-Scale_Vision_Transformer_for_Image_Classification_ICCV_2021_paper.html. Accessed April 12, 2023.

20. Hou R, Chang H, MA B, Shan S, Chen X. Cross Attention Network for Few-shot Classification. In: *Advances in Neural Information Processing Systems*.Vol 32. Curran Associates, Inc.; 2019. Available at: https://proceedings.neurips.cc/paper/2019/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html. Accessed April 12, 2023.

21. Hung ALY, Zheng H, Miao Q, Raman SS, Terzopoulos D, Sung K. CAT-Net: A Cross-Slice Attention Transformer Model for Prostate Zonal Segmentation in MRI. *IEEE Transactions on Medical Imaging* 2023;42:291–303.

22. Joo PY, Borjali A, Chen AF, Muratoglu OK, Varadarajan KM. Defining and predicting radiographic knee osteoarthritis progression: a systematic review of findings from the osteoarthritis initiative. *Knee Surg Sports Traumatol Arthrosc* 2022.

23. Tiulpin A, Melekhov I, Saarakkala S. KNEEL: Knee Anatomical Landmark Localization Using Hourglass Networks. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE; 2019:352–361. Available at: https://ieeexplore.ieee.org/document/9022083/. Accessed October 29, 2022.

24. Wang Y, Li S, Zhao B, Zhang J, Yang Y, Li B. A ResNet-based approach for accurate radiographic diagnosis of knee osteoarthritis. *CAAI Transactions on Intelligence Technology* 2022;7:512–521.

25. Liebel L, Körner M. Auxiliary Tasks in Multitask Learning. 2018. Available at: http://arxiv.org/abs/1805.06334. Accessed March 21, 2023.

26. Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, Meurs J van, et al. Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. *Sci Rep* 2019;9:20038.

27. Panfilov E, Tiulpin A, Nieminen MT, Saarakkala S. RADIOGRAPHIC OSTEOARTHRITIS PROGRESSION PREDICTION VIA MULTIMODAL IMAGING DATA AND DEEP LEARNING. *Osteoarthritis and Cartilage* 2022;30:S86–S87.

28. Yeh L-R, Zhang Y, Chen J-H, Liu Y-L, Wang A-C, Yang J-Y, et al. A deep learning-based method for the diagnosis of vertebral fractures on spine MRI: retrospective training and validation of ResNet. *Eur Spine J* 2022;31:2022–2030.

29. 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 Vis* 2020;128:336–359.

30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In: ; 2017:618–626. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html. Accessed April 12, 2023.

31. Demehri S, Kasaeian A, Roemer FW, Guermazi A. Osteoarthritis year in review 2022: imaging. *Osteoarthritis and Cartilage* 2023:S1063458423007264.

32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988;44:837–845.

33. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.

34. Kim H-E, Kim HH, Han B-K, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2020;2:e138–e148.

35. Hu K, Wu W, Li W, Simic M, Zomaya A, Wang Z. Adversarial Evolving Neural Network for Longitudinal Knee Osteoarthritis Prediction. *IEEE Transactions on Medical Imaging* 2022:1–1.

36. Han T, Kather JN, Pedersoli F, Zimmermann M, Keil S, Schulze-Hagen M, et al. Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nat Mach Intell* 2022;4:1029–1039.

37. Hirvasniemi J, Runhaar J, Heijden RA van der, Zokaeinikoo M, Yang M, Li X, et al. The KNee OsteoArthritis Prediction (KNOAP2020) challenge: An image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images. *Osteoarthritis and Cartilage* 2022:S1063458422008640.

38. Panfilov E, Saarakkala S, Nieminen MT, Tiulpin A. Predicting Knee Osteoarthritis Progression from Structural MRI using Deep Learning. 2022. Available at: http://arxiv.org/abs/2201.10849. Accessed January 10, 2023.

Table 1. Baseline Characteristics of Participants

	Training Set	**Validation Set**	**Testing Set 1**	**Testing Set 2**
Participant Characteristics	**N= 2227**	**N= 557**	**N=801**	**N= 2653**
Age (y)	60.9 ± 9.16	61.3 ± 9.33	60.3 ± 9.06	64.2 ± 9.00
Gender				
Male	894 (40.1%)	228 (40.9%)	302 (37.7%)	1080 (40.7%)
Female	1333 (59.9%)	329 (59.1%)	499 (62.3%)	1573 (59.3%)
BMI (kg/m^2)	28.2 ± 4.69	27.9 ± 4.46	29.2 ± 5.19	-
Enrolled site	B-D	B-D	A, E	A-E
Time point	Baseline	Baseline	Baseline	48-months
No. of knee readings	4454	1114	1602	5306
KLG				
0	1928 (43.3%)	469 (42.1%)	593 (37.0%)	2182 (41.1%)
1	839 (18.8%)	221 (19.8%)	276 (17.2%)	969 (18.3%)
2	1124 (25.2%)	282 (25.3%)	538 (33.6%)	1456 (27.4%)
3	563 (12.6%)	142 (12.7%)	195 (12.2%)	699 (13.2%)
TKA				
- No	4408 (99.0%)	1103 (99.0%)	1580 (98.6%)	5213 (98.2%)
- Yes	46 (1.0%)	11 (1.0%)	22 (1.4%)	93 (1.8%)
OA Progression				
- No	3837 (86.1%)	991 (89.0%)	1378 (86.0%)	4924 (92.8%)
- Yes	617 (13.9%)	123 (11.0%)	224 (14.0%)	382 (7.2%)

Mean data are ± standard deviation; data in parentheses are percentages
KLG: Kellgrene Lawrence grade; TKA: total knee arthroplasty

Table 2. Comparison of Prediction Performance of Bilateral Knee Neural Network and Unilateral Models

Model	**Testing set 1**			**Testing set 2**		
	AUC [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]	AUC [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]
Panfilov et al. benchmark[†]	0.71 (0.02)	-	-	-	-	-
ResNext	0.707 [0.670-0.743]	0.746 [0.688-0.799]	0.556 [0.53-0.583]	0.667 [0.640-0.693]	0.788 [0.746-0.830]	0.481 [0.467-0.495]
DenseNet	0.708 [0.669-0.744]	0.518 [0.451-0.580]	0.805 [0.784-0.824]	0.649 [0.621-0.677]	0.702 [0.654-0.746]	0.521 [0.507-0.536]
BikNet	**0.761**[*] **[0.728-0.795]**	0.665 [0.603-0.728]	0.774 [0.753-0.797]	**0.746**[*] **[0.721-0.768]**	0.675 [0.631-0.720]	0.738 [0.726-0.750]

[†] Their model only tested on the baseline

[*] DeLong test showed all P values < 0.001

Figure Legends

Figure 1. Schematic overview of the deep learning model for osteoarthritis (OA) progression prediction on bilateral knee radiographs. Firstly, a pre-trained Hourglass network was utilized to detect and segment the right and left knee from the radiograph. In this step, the radiograph was resized to 700 × 700 pixels. Subsequently, the cropped knee image was preprocessed to 310 × 310 pixels and utilized as the input for Bilateral Knee Neural Network (BikNet). BikNet was trained using a multitask deep learning approach for clinical diagnosis process simulation. Under the bilateral hypothesis, the auxiliary view will be input into cross-attention together with the main view to build up the cross-view information mappings. Finally, reader tests were conducted to evaluate the performance of the model in assisting in the diagnosis of early-stage OA.

Figure 2. Bilateral Knee Neural Network architecture. The left part of the figure shows that both the main and auxiliary views will undergo feature extraction through a backbone network and Attention mechanism. The Attention mechanism can help the model focus on the key structure of the knee rather than the unrelated image background. The feature from the main view is then used for auxiliary tasks to simulate clinical diagnosis for prediction reasoning. Afterwards, the cross-attention module will construct information bridges between the main view and auxiliary view to map unilateral features into bilateral features, which is later combined with the main view to predict the final OA progression status.

Figure 3. Performance of models and clinicians in OA progression prediction and early-stage OA diagnosis. A-B, comparison of model performance based on the areas under the ROC curves for (**A**) testing set 1 and (**B**) testing set 2. **C**, the average performance of all clinicians, represented by a dot (without model support) and a star (with model support). The black arrow indicates the increased sensitivity and specificity achieved by working with the model. **D**, a magnified region of the dashed rectangular area of the ROC curve (as outlined in **C**), with individual clinicians represented by open shapes (without model support) and filled shapes (with model support). The integration of our system can enhance the diagnostic performance of clinicians, as depicted by the dashed connection lines.

Figure 4. Visualization of representative cases of progression and non-progression, highlighting the focus of the Bilateral Knee Neural Network. The top column displays the original images, while the bottom column displays the Grad-CAMs. **A**, correctly predicted progression cases. **B**, correctly predicted non-progression cases. **C**, cases of incorrect prediction. Grad-CAM, gradient-weighted class activation map.

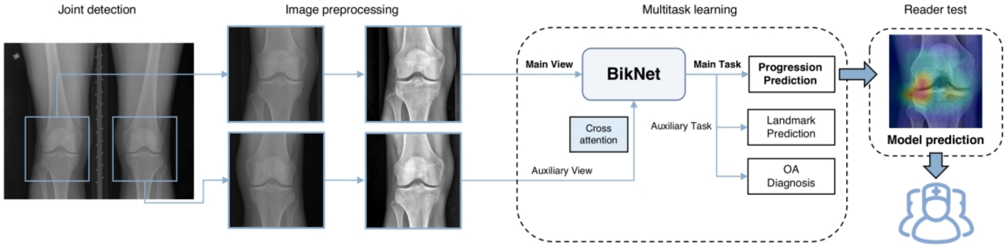


fig 1. schematic overview

289x75mm (300 x 300 DPI)

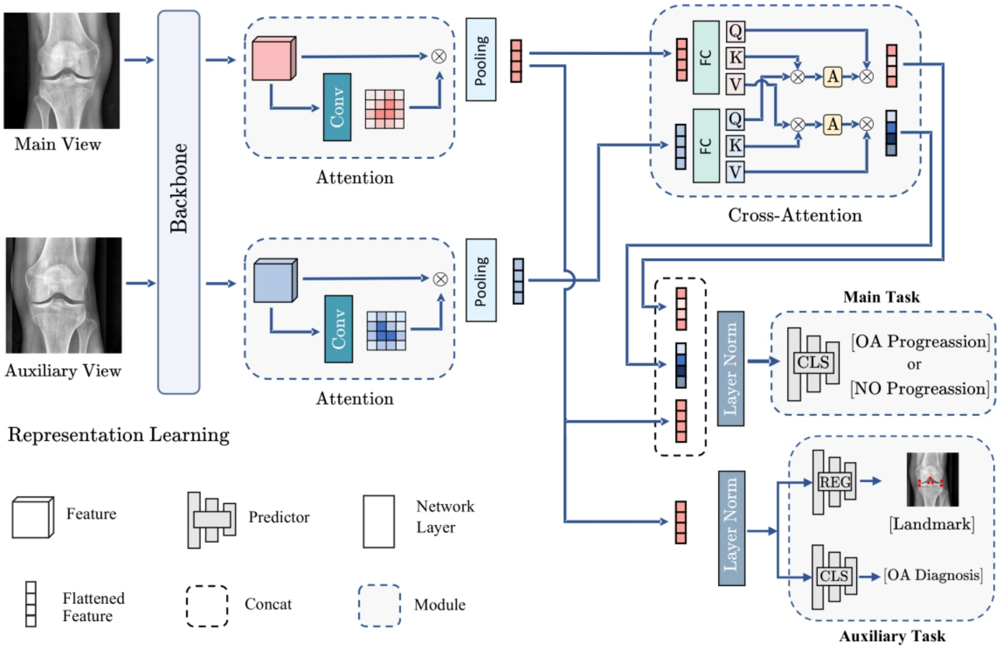


fig 2. model architecture

227x148mm (300 x 300 DPI)

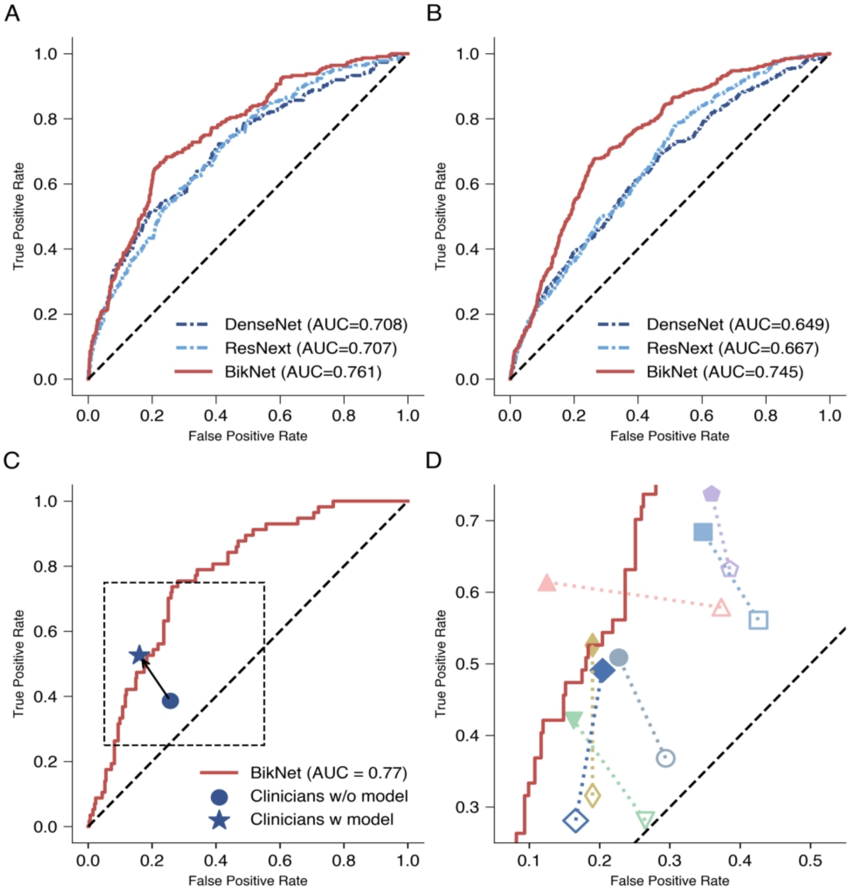


fig 3. model performance

228x211mm (300 x 300 DPI)

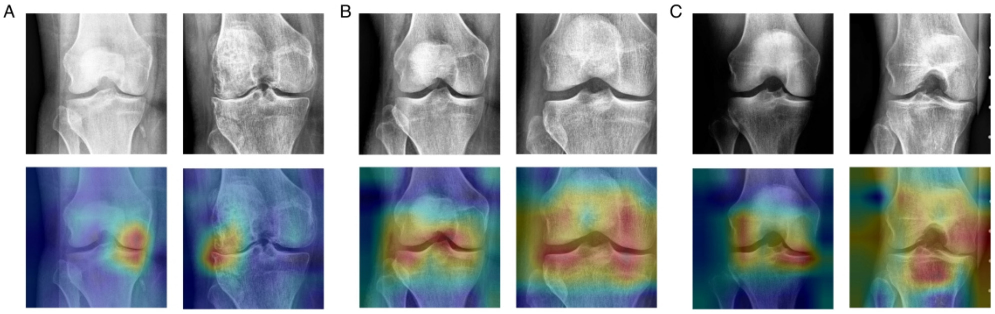


fig 4. grad-cam

280x88mm (300 x 300 DPI)

Appendix E1

Bilateral Hypothesis Justification. Consider a family $\mathcal{Y}$ of OA distributions, where $Y_n \in \mathcal{Y}$ is indexed by time stamp n ∈ $\mathrm{R}^{\mathrm{T}}_{\geq 0}$. Under unilateral setting, let A_n denotes main view in time n, we want to predict the OA condition Y_{n+1} in the future point $n+1$. To facilitate the expression, we will omit the subscript in the variables within the following formulas, e.g., the mutual information between variables Y_{n+1} and A_n is using $I(A;Y)$ instead of $I(A_n;Y_{n+1})$.

The $I(\cdot)$ indicates mutual information (MI), a larger MI value indicates greater variable relevance. While in bilateral hypothesis, we are given an additional auxiliary view B_n, where the mutual information can be denoted as:

$$I(A,B;Y) = I(A;Y) + I(B;Y \mid A).\#(1)$$

Where $I(A,B;Y)$ is the information A, B together provide about Y, and $I(B;Y \mid A)$ is the conditional mutual information between B and Y given A. Noted that, given $I(\cdot) \geq 0$, we have

$$I(A,B;Y) \geq I(A;Y),\#(2)$$

which indicates the we can have information gain $H = I(A,B;Y) - I(A;Y) = I(B;Y \mid A) \geq 0,$ of through providing model with Bilateral inputs. Therefore, this conclude that the model can enjoy performance benefit when providing bilateral input whereas afore-mentioned information gain is positive.

Model architecture. As shown in Fig. 2, BikNet consists of three modules: feature extraction, feature cross-fusion and multi-tasking modules. We employ the same feature extraction modules for either main view and auxiliary view, but use different feature fusion and forward processes for different views.

Given a main view input x_{main} and auxiliary view input x_{aux}, we use a backbone network $\mathcal{F}(\cdot)$ and a attention mechanism $\mathcal{A}_m(\cdot)$ as feature extraction modules:

$$z_{\text{main}} = \mathcal{A}_m(\mathcal{F}(x_{\text{main}})),\#(3)$$

$$z_{\text{aux}} = \mathcal{A}_m(\mathcal{F}(x_{\text{aux}})).\#(4)$$

In the context of bilateral hypothesis, the main view carries more relevant information in some patient cases, whereas the auxiliary view might be more relevant for others. Since multiple views convey diversified information combination possibility, their relationship needs to be effectively captured and connected. To be relieved from this complexity, we employ a cross-attention module to integrate cross-view information, which has a fully-connection layer FC to mapping feature into cross-attention space:

$$(Q_m,K_m,V_m) = \mathrm{FC}(z_{\mathrm{main}}),\ (Q_{\mathrm{a}},K_{\mathrm{a}},V_{\mathrm{a}}) = \mathrm{FC}(z_{\mathrm{aux}}),\#(5)$$

where Q stands for query, and (K,V) is the key-value vectors. The cross-attention $C(\cdot)$can calculated by:

$$\mathrm{u_m} = \mathrm{C}(\mathrm{Q_m},K_a,V_a) \ = \mathrm{softmax}(Q_m K_a)V_a,\#(6)$$

$$\mathrm{u_a} = \mathrm{C}(Q_a,K_m,V_m) \ = \mathrm{softmax}(Q_a K_m)V_m.\#(7)$$

The u is the cross-attentive feature. For example, the Eq. 6. represents for a specific query Q from main view, we use it to search within (K,V) from auxiliary view and output relevant information.

Finally, the above features are used in OA classification head and landmark regression head as shown in Fig. 2.

Training details. We utilized PyTorch, a popular open-source framework, to implement all deep learning models. The batch size was set to 32 and the AdamW optimizer was employed. The learning rate was warmed up linearly for 5 epochs from 2×10^{-4} to 2×10^{-3} and then set to 2×10^{-3} for the rest of the training. To address the issue of class imbalance, we oversampled the progression cases and applied focal loss (1) during training. The ConvNeXt (2) model was chosen as the backbone of our BikNet, and all models were pre-trained in ImageNet (3). The training set was augmented by applying random cropping, random rotation, horizontal and vertical flip, and gamma noise. Models were trained for up to 25 epochs, and the best snapshots were selected based on the area under the curve (AUC) at validation.

1. Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. In: ; 2017:2980–2988. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html. Accessed April 11, 2023.

2. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. 2022. Available at: http://arxiv.org/abs/2201.03545. Accessed April 11, 2023.

3. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*.; 2009:248–255.

Supplement Tables

Table S1. Comparison of Prediction Performance of Bilateral Knee Neural Network and other Unilateral models

Model	Testing set 1			Testing set 2		
	AUC [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]	AUC [95% CI]	Sensitivity [95% CI]	Specificity [95% CI]
EfficientNet	0.655 [0.616-0.696]	0.576 [0.509-0.643]	0.687 [0.661-0.710]	0.652 [0.625-0.677]	0.733 [0.688-0.777]	0.510 [0.497-0.523]
ResNet34	0.681 [0.643-0.716]	0.652 [0.585-0.710]	0.622 [0.596-0.647]	0.654 [0.626-0.682]	0.586 [0.537-0.636]	0.665 [0.653-0.678]
ResNet50	0.669 [0.631-0.709]	0.536 [0.469-0.598]	0.758 [0.734-0.780]	0.646 [0.618-0.673]	0.681 [0.634-0.725]	0.549 [0.535-0.563]
BikNet	**0.761*** **[0.728-0.795]**	0.665 [0.603-0.728]	0.774 [0.753-0.797]	**0.746*** **[0.721-0.768]**	0.675 [0.631-0.720]	0.738 [0.726-0.750]

* DeLong test showed all P value < 0.001

Table S2. Performance of Individual Clinicians in Predicting OA Onset with or without the Assistance of Model

Clinicians (Years of experience)	Model assistance	Sensitivity (%)	Specificity (%)
Orthopedist 1	No	36.8	70.6
(5 years)	Yes	50.9	77.3
Orthopedist 2	No	28.1	73.5
(4 years)	Yes	42.1	83.7
Orthopedist 3	No	56.1	57.4
(11 years)	Yes	68.4	65.3
Orthopedist 4	No	63.2	61.5
(24 years)	Yes	73.7	64.1
Radiologist 1	No	31.6	81.0
(6 years)	Yes	52.6	81.0
Radiologist 2	No	28.1	83.4
(4 years)	Yes	49.1	79.6
Radiologist 3	No	57.9	62.7
(12 years)	Yes	61.4	87.5

Table S3. Average Ratings of Clinicians in Predicting OA Onset

	Sensitivity	Specificity	Fleiss' Kappa
Model assistance			
Without model	0.386 [0.248-0.511]	0.743 [0.696-0.791]	0.203
With model	0.526 [0.395-0.664]	0.840 [0.801-0.881]	0.365

Supplement Figure Legends

Figure S1. Flowchart showing participant selection and datasets formation from the Osteoarthritis Initiative (OAI).

Figure S2. Overview of the image preprocessing pipeline. The cropped X-ray image is first re-oriented so that both left and right knees are similarly oriented. The histogram clipping is then applied, followed by a histogram normalization for trabecular texture enhancement.

Figure S3. Images show an application of the customized program used by clinicians to diagnose "early-stage" OA. A, prediction without model support. **B**, prediction with model support

Figure S4. Comparison performances among BikNet and other commonly used deep learning models based on the areas under the ROC curves. A, testing set 1. **B**, testing set 2.

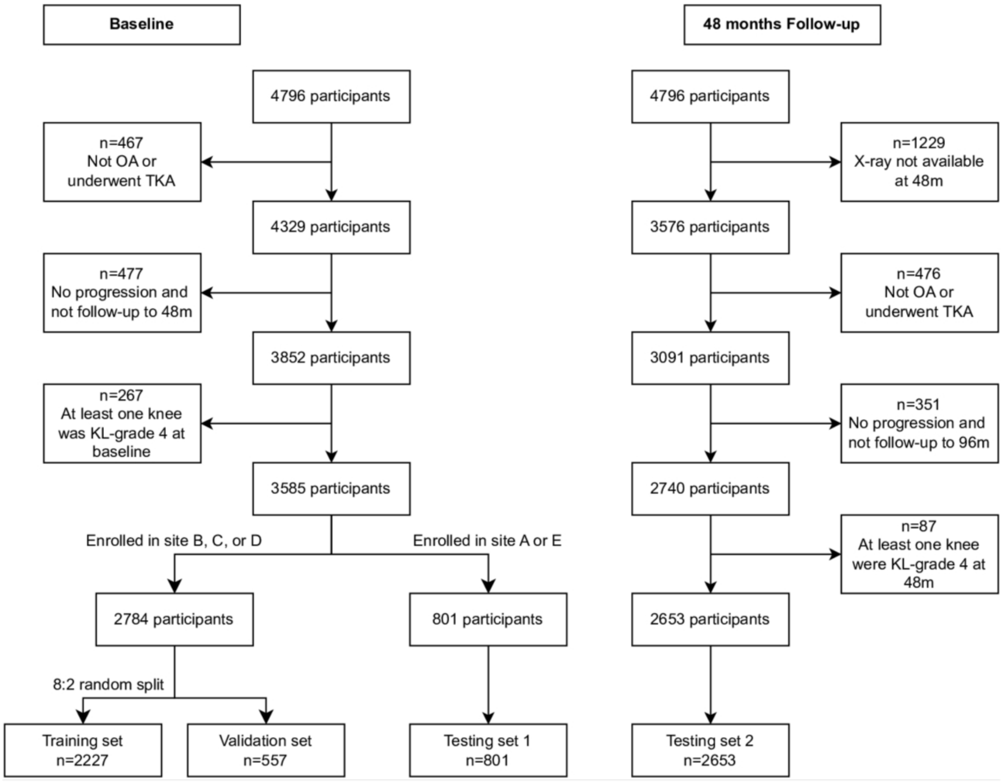


fig s1. flowchart

227x177mm (300 x 300 DPI)

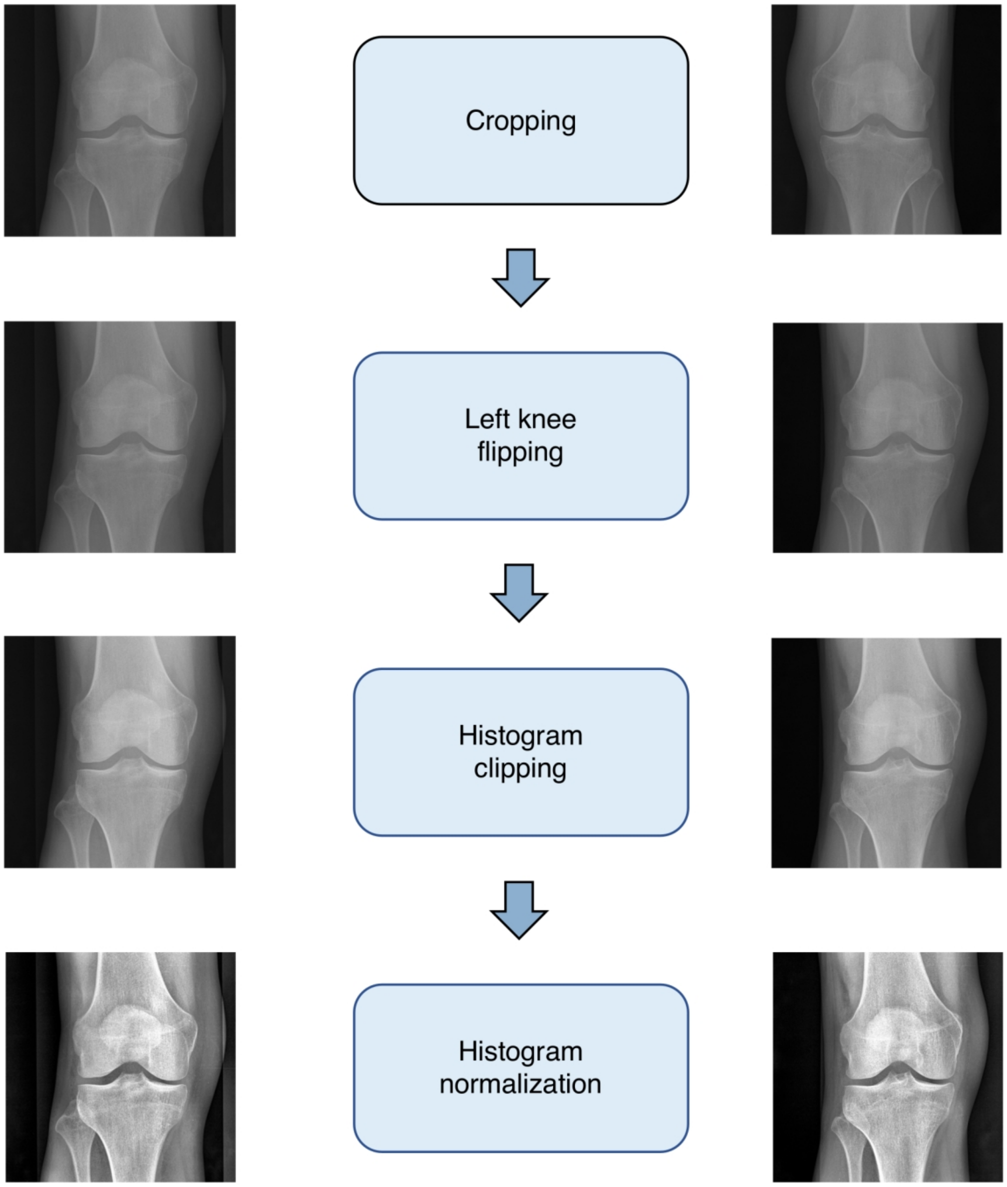
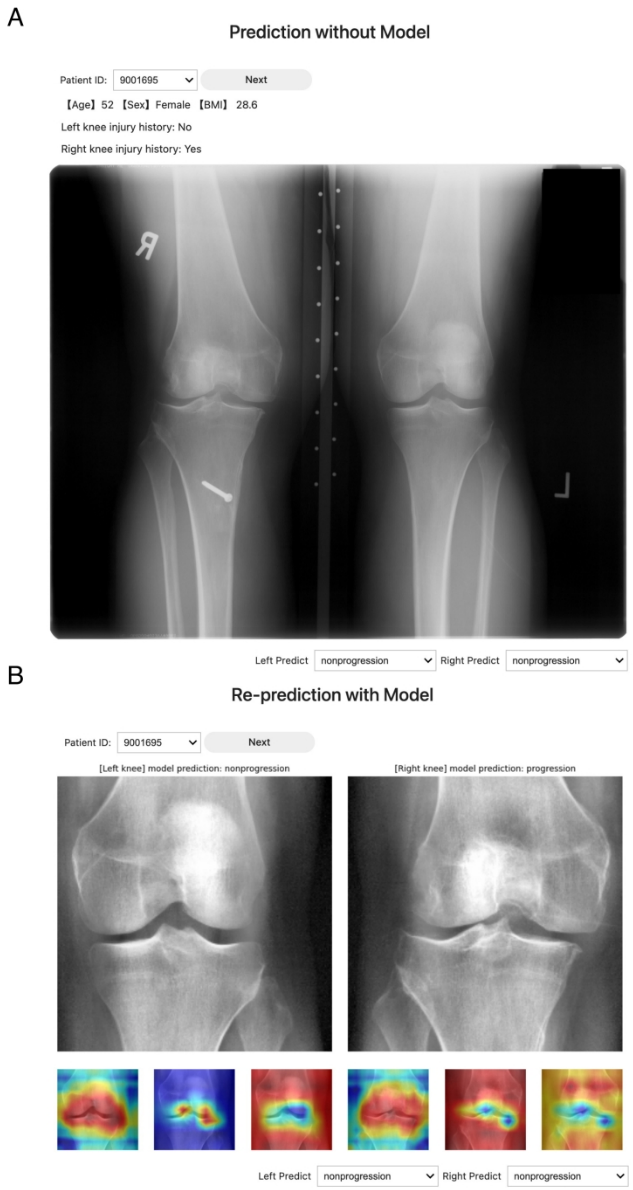


fig s2. image preprocessing

151x178mm (300 x 300 DPI)



Prediction without Model

Patient ID: 9001695

【Age】52 【Sex】Female 【BMI】 28.6

Left knee injury history: No

Right knee injury history: Yes

Left Predict nonprogression Right Predict nonprogression

Re-prediction with Model

Patient ID: 9001695

[Left knee] model prediction: nonprogression

[Right knee] model prediction: progression

Left Predict nonprogression Right Predict nonprogression

fig s3. reader test interface

149x273mm (300 x 300 DPI)

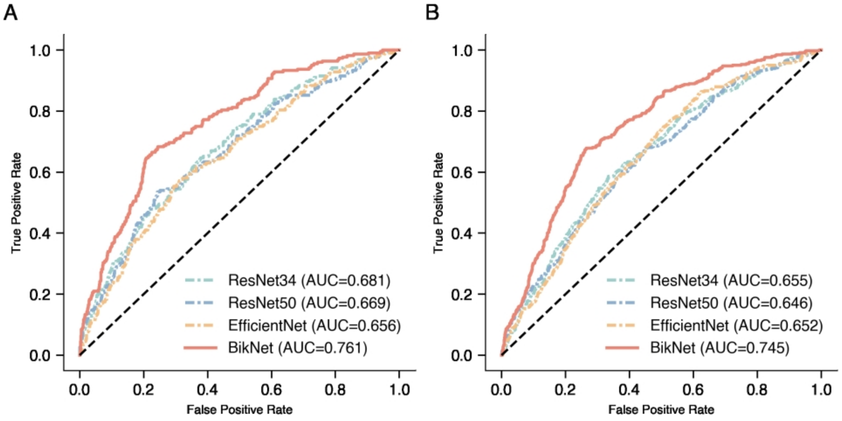


fig s4. supplement model performance

228x101mm (300 x 300 DPI)