Osteoarthritis and Cartilage 32 (2024) 338–347

Osteoarthritis and Cartilage

Expanding from unilateral to bilateral: A robust deep learning-based approach for predicting radiographic osteoarthritis progression

Rui Yin # †, Hao Chen ‡, Tianqi Tao †, Kaibin Zhang †, Guangxu Yang §, Fajian Shi §, Yiqiu Jiang # †, Jianchao Gui # † *

Nanjing Medical University, Nanjing, China
† *Department of Sports Medicine and Joint Surgery, Nanjing First Hospital, Nanjing, China*
‡ *School of Computer Science, University of Birmingham, Birmingham, UK*
§ *Department of Orthopedic Surgery, Nanjing Pukou Hospital, Nanjing, China*

ARTICLE INFO

Article history:
Received 24 April 2023
Accepted 29 November 2023

Keywords:
Osteoarthritis
Deep learning
Progression
Cross attention
X-ray

SUMMARY

Objective: To develop and validate a deep learning (DL) 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 predicting incident OA.
Results: Totally 3583 participants were evaluated. The BikNet we proposed outperformed ResNext and DenseNet (all area under the curve [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 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 incident OA prediction and improved inter-observer reliability.
Conclusion: The DL model, constructed based on bilateral knee views, holds promise for enhancing the assessment of OA and demonstrates greater robustness during subsequent follow-up evaluations as compared with unilateral models. 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.[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. 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,

Correspondence to: Department of Sports Medicine and Joint Surgery, Nanjing First Hospital, Nanjing Medical University, Changle Road 68, Nanjing 210006, China.
E-mail addresses: ray_yin@foxmail.com (R. Yin), h.chen.12@bham.ac.uk (H. Chen), 18360862922@139.com (T. Tao), kaibin_zhang09@163.com (K. Zhang), yangguangxu1980@163.com (G. Yang), jbgk@sina.com (F. Shi), jyq_3000@163.com (Y. Jiang), gui1997@126.com (J. Gui).

https://doi.org/10.1016/j.joca.2023.11.022

circumventing the need for laborious manual feature engineering and surpassing the performance of conventional 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, most previous works were just fine-tuning pre-trained general DL models, which 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, an OA-specific model need to account for both knees concurrently when assessing the relationship between symptoms, physical function, and structural disease, as clinicians 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 this study, we proposed the Bilateral Knee Neural Network (BikNet) as an OA-specific architecture to address the limitations of previous DL models for OA. BikNet incorporates a cross-attention module[19–21] and multi-task learning,[22] enabling simultaneous evaluation of both knees and capturing their interdependence. By leveraging information from bilateral views in raw X-ray images, BikNet aims to provide more accurate predictions. Our hypothesis is that BikNet can learn more effective representations from the contralateral joint, outperforming previous DL models (unilateral models) that assess one knee at a time, as well as simple bilateral version models in predicting OA progression at baseline and subsequent follow-up time points. Furthermore, we contend that BikNet can aid in predicting OA onset.

Methods

Datasets

This retrospective study analyzed radiographs from a total of 12,650 knees, obtained from 3585 participants enrolled in 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 1211 participants were excluded for various reasons, including unavailability of Kellgren-Lawrence grade (KLG), knee replacement surgery, diagnosed inflammatory arthritis, at least one knee with KLG 4, or follow-up duration of less than 48 months without confirmed OA progression (Fig. 1). Baseline radiographs (n = 3585) 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 = 2227) and 20% (n = 557), respectively. To further evaluate the models' robustness and mimic clinical scenarios, testing set 2 (n = 2653) 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 at each site. All individuals gave written informed consent. The bilateral standing posterior-anterior knee X-rays were acquired using a standardized technique, employing a SynaFlexor frame[23] and a 10-degree beam angle.

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.[24]

DL workflow

The DL workflow is depicted in Fig. 2. 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 DL 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 Fig. S1. First, a pre-trained Hourglass network[25] was employed to extract a region of interest of size 700 × 700 pixels from each knee in the bilateral posteroanterior fixed-flexion knee radiographs. To ensure the inclusion of valid joint information, the cropped images underwent quality control conducted by two radiologists. Any cases where the center of the joint deviated from the central 350 × 350 pixel region were excluded instead of opting for manual cropping. This approach guaranteed an automated preprocessing process, enhancing efficiency and consistency. Then the left knee images were flipped to the right knee configuration and resized to 310 × 310 pixels. 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.[26]

Model architecture

The diagram of our model's architecture is illustrated in Fig. 3. 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.

Meanwhile, a multitask learning paradigm was employed 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.[22] 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. 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 our model architecture, we compared it with the best-performing backbones (DenseNet and ResNext)

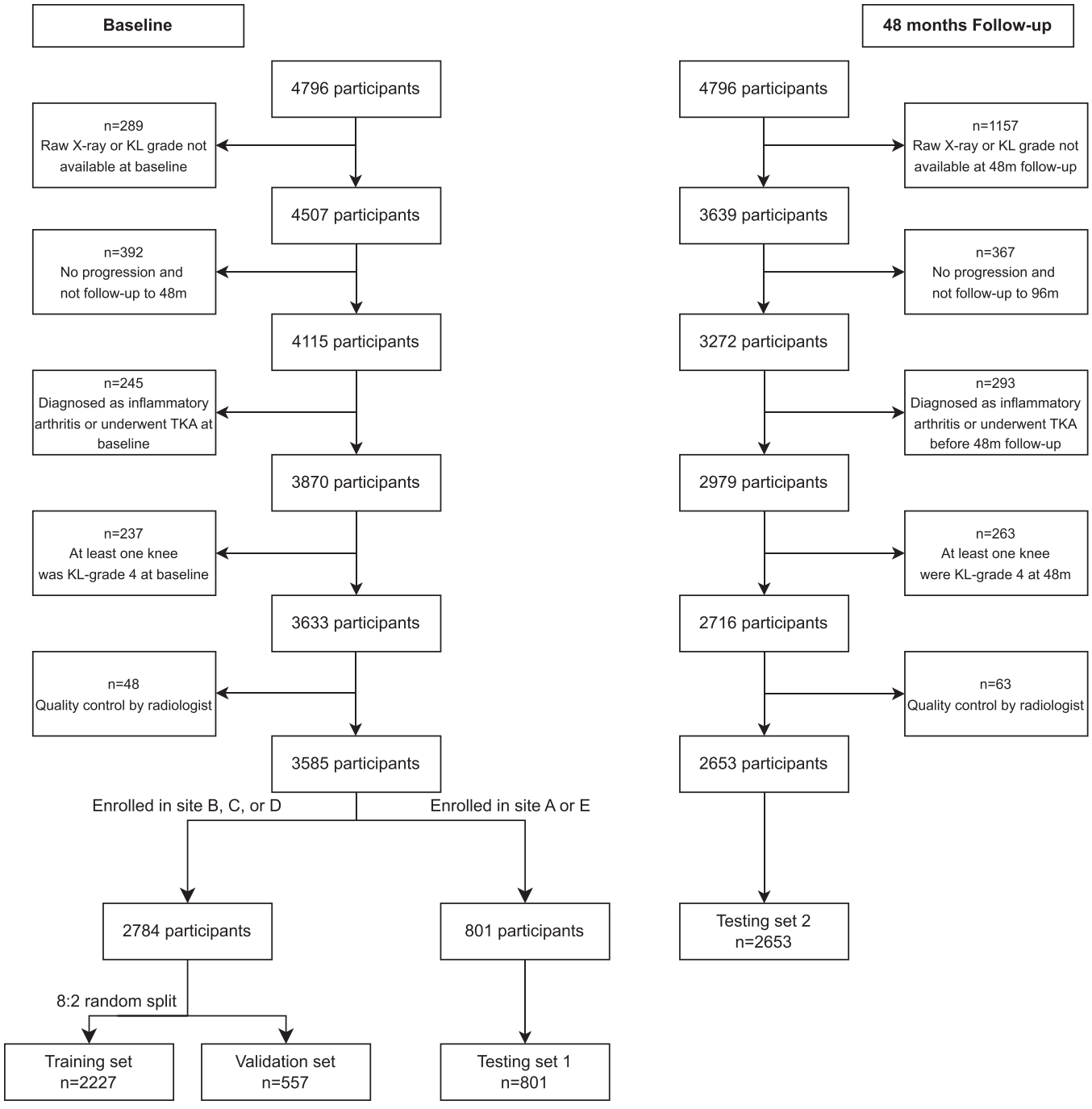

Fig. 1 Osteoarthritis and Cartilage

Flowchart showing participants selection and dataset formation from the OAI.

from previous studies that predicted OA progression, which served as unilateral convolutional neural network models.[14,27] The result reported by Panfilov et al.[28] 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. To ensure a fair and comprehensive comparison, we also trained the bilateral versions of DenseNet and ResNext, referred to as BiDenseNet and BiResNext, respectively. In this setup, both the left and right knee images shared the same backbone, and their outputs features were concatenated before the final classifier. Additionally, we evaluated other commonly used DL models in medical imaging, including ResNet34, ResNet50, and EfficientNet, to supplement our analysis.[12,15,26,29] 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

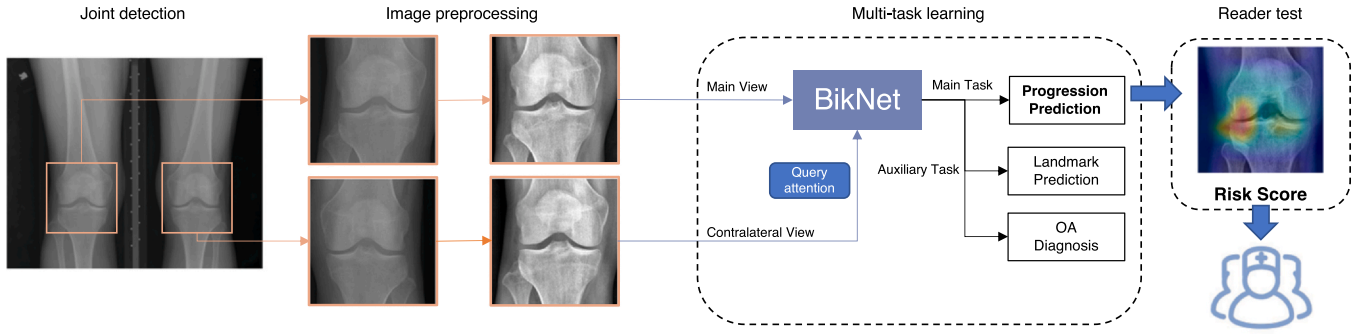

Fig. 2 Osteoarthritis and Cartilage

Schematic overview of the DL model for 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. During this process, the cropped image of each knee was resized to 700 × 700 pixels. Subsequently, the cropped knee image was preprocessed to 310 × 310 pixels and utilized as the input for BikNet. BikNet was trained using a multitask DL 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.

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.[30,31]

Reader test

Differentiating individuals with an impending onset of disease is crucial for identifying patients who require preventive care and has real potential to better define OA subgroups.[6,32] In this study, we defined incident OA as those without radiographic OA (KLG 0-1) at

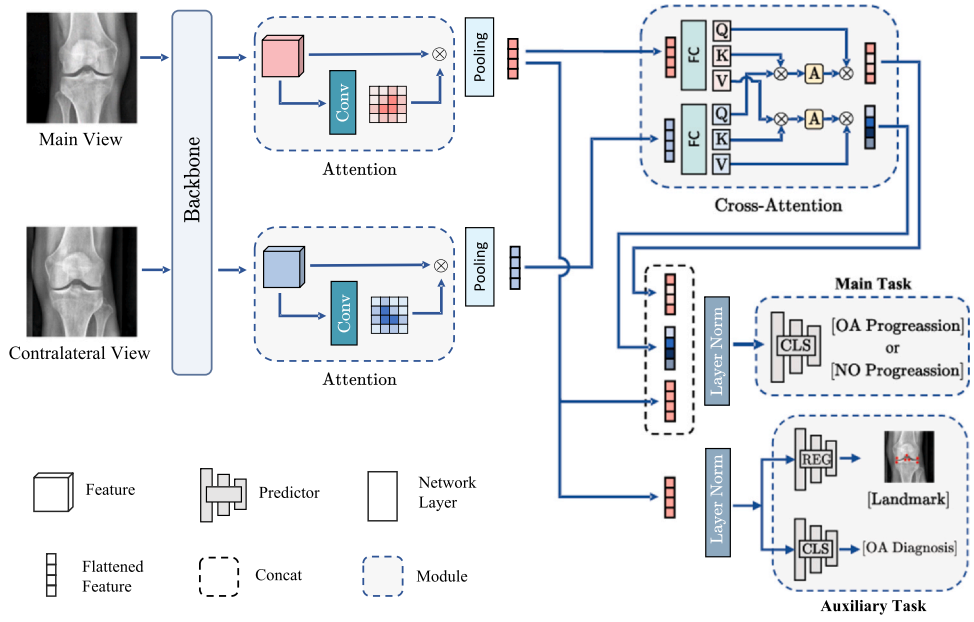

Fig. 3 Osteoarthritis and Cartilage

Bilateral Knee Neural Network architecture. The left part of the figure shows that both the main and contralateral 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 contralateral view to map unilateral features into bilateral features, which is later combined with the main view to predict the final OA progression status.

baseline but showed progression of one or more KLGs (KLG ≥2) over four years. Two experiments were conducted to evaluate the performance of our model in assisting with the prediction of incident OA. In Experiment 1, seven experienced clinicians, including four orthopedists and three radiologists, were given only bilateral 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 incident OA cases (57 among 400 knees), from two testing sets. All reading experiments were performed on diagnostic computer monitors. Fig. S2 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 other models, receiver operator characteristic (ROC) analysis was used, and the AUCs were calculated. Standard deviations and 95% confidence intervals were obtained through bootstrapping with 2000 redraws. The Youden index was used to determine optimal model sensitivity and specificity. To compare the AUCs of BikNet and other models, we employed the DeLong test.[33] Additionally, 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 of 28.3 ± 4.79 kg/m^2 at baseline. Among the 3583 participants, 2161 were women, which accounted for 59.3% of the sample. In the subsequent follow-up period (testing set 2), the mean age of the 2653 participants was 64.2 ± 9.00 years, with 1573 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 I 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 II presents the results of using Panfilov et al.[28] as the benchmark for our study, where 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 Fig. 4A and 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). Similarly, BikNet achieved the highest AUC in testing set 2 with a value of 0.746 [0.721-0.768], 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. It also significantly outperformed simple bilateral versions of models (all P < 0.001). Notably, simple concatenation of the contralateral image could not increase model performance. The bilateral versions of DenseNet and ResNext did not demonstrate superiority compared to their unilateral counterparts, except for BiDenseNet on testing set 2 (BiDenseNet vs. DenseNet: 0.678 vs. 0.649). This finding highlights the importance of an advanced architecture in BikNet, which incorporates specific cross-attention mechanisms to effectively leverage information from both knees and

Participant characteristics	Training set N = 2227	Validation set N = 557	Testing set 1 N = 801	Testing set 2 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.
BMI: body mass index.

Table I

Baseline characteristics of participants.

Osteoarthritis and Cartilage

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]
BiResNext‡	0.664 [0.624-0.704]	0.478 [0.406-0.545]	0.811 [0.79-0.832]	0.656 [0.627-0.684]	0.657 [0.61-0.704]	0.591 [0.577-0.604]
BiDenseNet‡	0.700 [0.663-0.737]	0.696 [0.638-0.754]	0.613 [0.587-0.638]	0.678 [0.651-0.705]	0.657 [0.61-0.707]	0.612 [0.599-0.626]
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.
‡Simple concatenation of bilateral view.
*DeLong test showed all P values < 0.001.
CI: confidence interval.
The purpose of bold is to highlight the performance of our model, which is significantly superior to other methods.

Table II

Comparison of prediction performance of Bilateral Knee Neural Network and other models.

optimize the predictive capability for OA progression. BikNet significantly outperformed other commonly used models 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 Fig. S3 and Table SI.

Assistance in the prediction of incident OA

To assess the effectiveness of our model in assisting clinicians with the prediction of incident OA, we conducted two reader tests. In the first experiment, most clinicians were unable to reliably differentiate between the two groups, except for one joint specialist (F.J.). 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 SII). This was expected as the current approach did not enable clinicians to predict incident OA. In the second test, results improved substantially with the additional informative presentation of the model predictions. As shown in Table SII, 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 (Fig. 4C and D). It was also noteworthy that model support helped 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 SIII).

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 Fig. 5, 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 (Fig. 5A), the model's attention was primarily on the medial/lateral joint space or osteophytes, while for nonprogression OA (Fig. 5B), 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. Fig. 5C presents examples of prediction errors, which could potentially be attributed to factors such as image degradation, artifacts, or obscured bony structures.

Ablation study

To assess the effectiveness of different components in our proposed architecture, ablation studies were conducted. The base model simply concatenated the two view features, similar to BiDenseNet and BiResNext. As shown in Table SIV, the inclusion of the cross-attention module resulted in a substantial increase in AUC of 10.34% and 10.66% in testing sets one and two, respectively, compared to the base model. Furthermore, the incorporation of the two auxiliary tasks yielded the best performance, leading to an additional 5% improvement in AUC. These findings demonstrate the effectiveness of both the cross-attention and the multi-task learning modules.

Discussion

Our study presents a fully automated DL-based system for predicting OA progression 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 and follow-up stages. Moreover, BikNet considerably enhanced the sensitivity and specificity of incident OA prediction by clinicians, highlighting the promising potential of computer-based methods for evaluating OA.

Although radiographic features have limited added value in predicting OA progression, previous studies have confirmed the potential of DL in assessing OA using radiographs. Guan et al.[14] utilized a DenseNet 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.[27] 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 previous methods. Panfilov et al.[28] extended Tiulpin's approach and reported an AUC of 0.71 for a DL method based on X-ray in predicting OA progression.

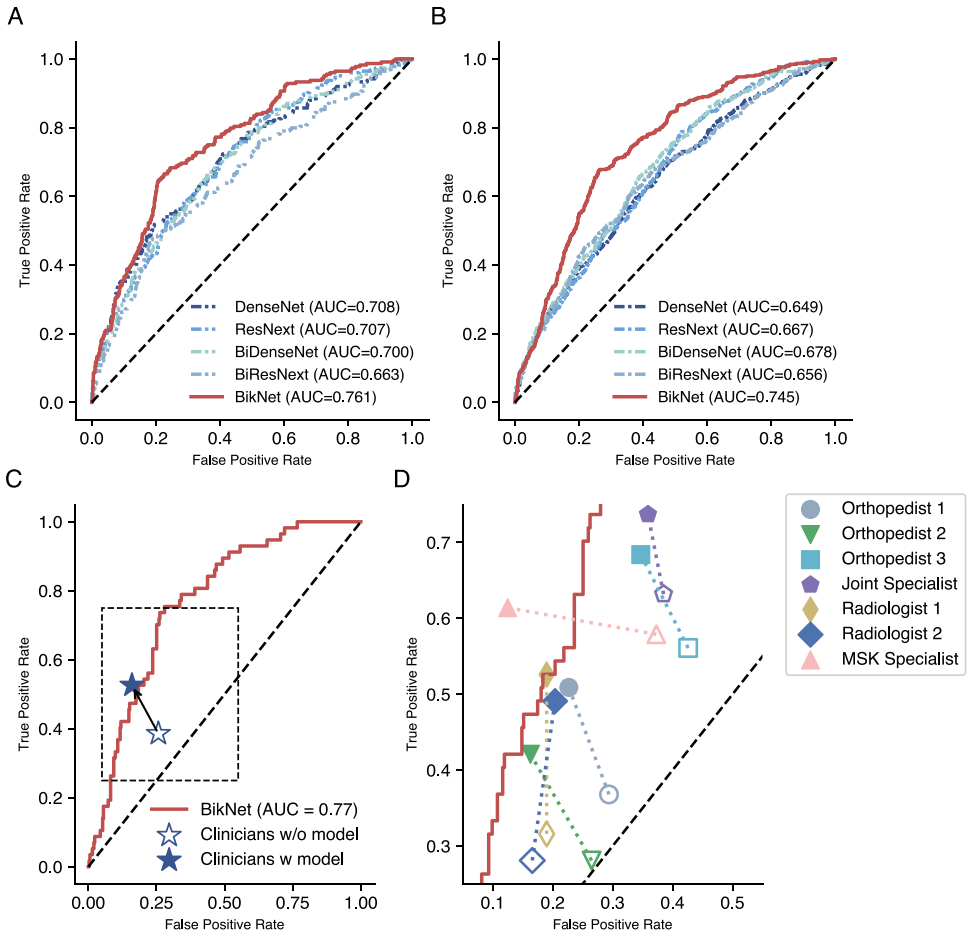

Fig. 4

Osteoarthritis and Cartilage

Performance of models and clinicians in predicting OA progression and incident OA cases. 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 an unfilled shape (without model support) and a filled shape (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.

However, prior studies on DL for OA have taken 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 architecture that can assess bilateral knees simultaneously.

Merely concatenating two views, however, may not lead to improved model performance; in fact, it could potentially result in the model learning irrelevant features, leading to overfitting. This issue was evident when comparing the bilateral versions of DenseNet and ResNext, as they did not outperform their unilateral versions. To address this concern, our proposed model incorporates a cross-attention module to effectively fuse the information from both knees. This mechanism allows the model to focus on the relevant and informative features while minimizing the impact of irrelevant or noisy features. The results of ablation study demonstrated that by utilizing the cross-attention, the model's performance significantly increased. To further enhance the learning of useful features, we drew inspiration from the clinical diagnostic process and introduced two auxiliary tasks: OA diagnosis and landmarks prediction. Through multi-task learning, the model's performance was further improved, leading to a final model with strong discriminative ability. Moreover, as a degenerative 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 models declined significantly and exhibited weak discrimination during follow-up. However, due to the effective fusion of the features from the contralateral view, BikNet maintained a fair discrimination ability.

Recent studies have shown the potential benefits of DL-aided systems for various clinical applications. For instance, McKinney et al.[34] 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.[35] Their results showed that the diagnostic accuracy

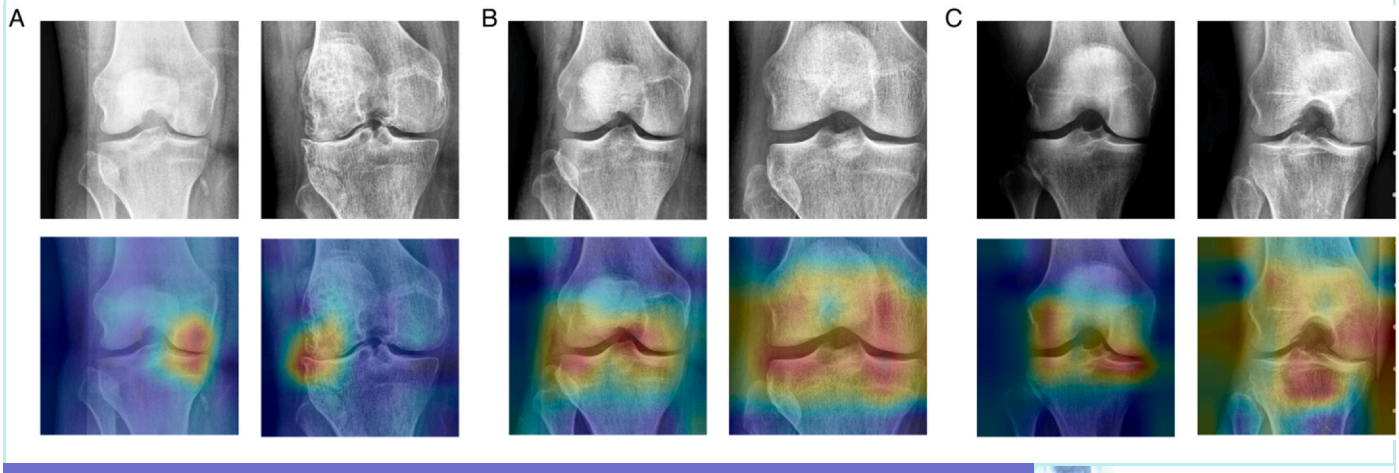

Fig. 5

Osteoarthritis and Cartilage

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.

of radiologists was significantly enhanced when working alongside 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 incident OA prediction. 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 (Fig. SIIIA). 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. 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 predict incident OA patients based only on clinical information and X-rays. This could facilitate early diagnosis and prompt intervention for OA in the future.

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 models that can handle variations in image quality and disease progression over time.[36,37] 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).[28,38,39] Despite 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, or other modality images, to aid in the assessment of OA.

Our study has several limitations. Firstly, the data utilized was obtained solely from the OAI, which has a limited representation of the Asian population.[5] Therefore, it is necessary to validate the efficacy of BikNet 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.[24] However, the difference in definition means that our model cannot be directly compared with some previous models.[14,27] Additionally, it is important to note that the OAI employs specialized X-ray protocols that may not be commonly used in clinical practice. Nevertheless, prior research has demonstrated that the ResNext model achieved good performance in other datasets, such as MOST, which utilized a more common protocol.[27] In the future, exploring generative models such as Generative Adversarial Networks[40] or Stable Diffusion[41] for style transfer between different protocols could potentially overcome this issue, enhancing the clinical utility of OA prediction models.

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 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.

Role of funding sources

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

Author contribution

Rui Yin: Conception of design, analysis and interpretation of data, model development, and drafting of the article. Hao Chen: Model development and optimization. Tianqi Tao: Reader test. Kaibin Zhang: Reader test. Guangxu Yang: Clinical expertise, reader test. Fajian Shi: Clinical expertise, reader test. Yiqiu Jiang: Clinical expertise, image quality control. Jianchao Gui: Conception of design, clinical and statistical expertise. All authors contributed to the drafting of the article and final approval of the version to be submitted.

Conflict of interest

The authors declare no competing interests.

Acknowledgments

We gratefully acknowledge data collection and curation efforts from the Osteoarthritis Initiative.

We acknowledge our colleagues in the radiology department for their help with the reader test.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.joca.2023.11.022.

References

1. Hunter DJ, Bierma-Zeinstra S. Osteoarthritis. Lancet 2019;393(10182):1745–59. https://doi.org/10.1016/S0140-6736(19)30417-9
2. Sharma L. Osteoarthritis of the knee. N Engl J Med 2021;384(1):51–9. https://doi.org/10.1056/NEJMcp1903768
3. Yazici Y, McAlindon TE, Gibofsky A, Lane NE, Clauw D, Jones M, *et al*. Lorecivivint, a novel intraarticular CDC-like kinase 2 and dual-specificity tyrosine phosphorylation-regulated kinase 1A inhibitor and Wnt pathway modulator for the treatment of knee osteoarthritis: A phase II randomized trial. Arthritis Rheumatol 2020;72(10):1694–706. https://doi.org/10.1002/art.41315
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. Ann Rheum Dis 2021;80(8):1062–9. https://doi.org/10.1136/annrheumdis-2020-219181
5. Driban JB, Harkey MS, Barbe MF, *et al*. Risk factors and the natural history of accelerated knee osteoarthritis: a narrative review. BMC Musculoskelet Disord 2020;21(1), 332. https://doi.org/10.1186/s12891-020-03367-2
6. Foster NE, Eriksson L, Deveza L, Hall M. Osteoarthritis year in review 2022: epidemiology & therapy. Osteoarthr Cartil 2023;31(7):876–83. https://doi.org/10.1016/j.joca.2023.03.008
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 2020;60(5):2448–55. https://doi.org/10.1093/rheumatology/keaa643
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. Semin Arthritis Rheum 2022;55, 152007. https://doi.org/10.1016/j.semarthrit.2022.152007
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. https://doi.org/10.1016/j.media.2022.102444
10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, *et al*. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005
11. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28(9):1773–84. https://doi.org/10.1038/s41591-022-01981-2
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(2):363–73. https://doi.org/10.1007/s00256-021-03773-0
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(1):136–40. https://doi.org/10.1038/s41591-020-01192-7
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. Osteoarthr Cartil 2020;28(4):428–37. https://doi.org/10.1016/j.joca.2020.01.010
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(3):584–93. https://doi.org/10.1148/radiol.2020192091
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. Osteoarthr Cartil 2016;24(5):807–13. https://doi.org/10.1016/j.joca.2015.12.005
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. Osteoarthr Cartil 2013;21:S58–9. https://doi.org/10.1016/j.joca.2013.02.132
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(1), 153. https://doi.org/10.1186/1471-2474-13-153
19. Chen CF (Richard), Fan Q, Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification. 2021:357-366. Accessed April 12, 2023. https://openaccess.thecvf.com/content/ICCV2021/html/Chen_CrossViT_Cross-Attention_Multi-Scale_Vision_Transformer_for_Image_Classification_ICCV_2021_paper.html
.
20. Hou R, Chang H, Shan MAB, Chen S, Cross X. Attention network for few-shot classification. Adv Neural Inf Process Syst Vol 32. Curran Associates, Inc.; 2019 Accessed April 12, 2023 ⟨https://proceedings.neurips.cc/paper/2019/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html⟩.

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 Trans Med Imaging 2023;42(1):291–303. https://doi.org/10.1109/TMI.2022.3211764
22. Liebel L, Körner M. Auxiliary tasks in multi-task learning. Published online May 17, 2018. Accessed March 21, 2023 ⟨http://arxiv.org/abs/1805.06334⟩.
23. Kothari M, Guermazi A, Ingersleben G von, Miaux Y, Sieffert M, Block JE, *et al*. Fixed-flexion radiography of the knee provides reproducible joint space width measurements in osteoarthritis. Eur Radiol 2004;14:1568–73.
24. 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;7(3):512–21. https://doi.org/10.1007/s00167-021-06768-5
25. Tiulpin A, Melekhov I, Saarakkala S. KNEEL: Knee anatomical landmark localization using hourglass networks. 2019 IEEE/CVF International Conference on Computer Vision Workshop ((ICCVW)). Piscatawa, NJ: IEEE; 2019. p. 352–61. https://doi.org/10.1109/ICCVW.2019.00046
26. Wang Y, Li S, Zhao B, Zhang J, Yang Y, Li B. A ResNet-based approach for accurate radiographic diagnosis of knee osteoarthritis. CAAI Trans Intell Technol 2022;7(3):512–21. https://doi.org/10.1049/cit2.12079
27. 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(1), 20038. https://doi.org/10.1038/s41598-019-56527-3
28. Panfilov E, Tiulpin A, Nieminen MT, Saarakkala S. Radiographic osteoarthritis progression prediction via multi-modal imaging data and deep learning. Osteoarthr Cartil 2022;30:S86–7. https://doi.org/10.1016/j.joca.2022.02.107
29. 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(8):2022–30. https://doi.org/10.1007/s00586-022-07121-1
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128(2):336–59. https://doi.org/10.1007/s11263-019-01228-7
31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017:618–626. Accessed April 12, 2023 ⟨https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html⟩.
32. Demehri S, Kasaeian A, Roemer FW, Guermazi A. Osteoarthritis year in review 2022: imaging. Osteoarthr Cartil 2023;31(8):1003–11. https://doi.org/10.1016/j.joca.2023.03.005
33. 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(3):837–45. https://doi.org/10.2307/2531595
34. 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(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6
35. 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. Lancet Digital Health 2020;2(3):e138–48. https://doi.org/10.1016/S2589-7500(20)30003-0
36. Hu K, Wu W, Li W, Simic M, Zomaya A, Wang Z. Adversarial evolving neural network for longitudinal knee osteoarthritis prediction. IEEE Trans Med Imaging 2022;41(11):3207–17. https://doi.org/10.1109/TMI.2022.3181060
37. 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(11):1029–39. https://doi.org/10.1038/s42256-022-00560-x
38. Hirvasniemi, Runhaar J, Heijden J, van der RA, 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. Osteoarthr Cartil 2022;31(1):115–25. https://doi.org/10.1016/j.joca.2022.10.001
39. Panfilov E, Saarakkala S, Nieminen MT, Tiulpin A. Predicting knee osteoarthritis progression from structural MRI using deep learning. Published online January 26, 2022. Accessed January 10, 2023 ⟨http://arxiv.org/abs/2201.10849⟩.
40. Tolkach Y, Wolgast LM, Damanakis A, Pryalukhin A, Schallenberg S, Hulla W, *et al*. Artificial intelligence for tumour tissue detection and histological regression grading in oesophageal adenocarcinomas: a retrospective algorithm development and validation study. Lancet Digital Health 2023;5(5):e265–75. https://doi.org/10.1016/S2589-7500(23)00027-4
41. Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, Hacihaliloglu I, et al. Diffusion models for medical image analysis: A comprehensive survey. Published online November 14, 2022. Accessed December 30, 2022. http://arxiv.org/abs/2211.07804.