

<https://doi.org/10.1038/s41746-024-01265-8>

# Identifying who are unlikely to benefit from total knee arthroplasty using machine learning models

Check for updates

Xiaodi Liu <sup>1,4</sup>✉, Yingnan Liu<sup>1,2,4</sup>, Mong Li Lee<sup>1,2</sup>, Wynne Hsu<sup>1,2</sup> & Ming Han Lincoln Liow <sup>3</sup>

Identifying and preventing patients who are not likely to benefit long-term from total knee arthroplasty (TKA) would decrease healthcare expenditure significantly. We trained machine learning (ML) models (image-only, clinical-data only, and multimodal) among 5720 knee OA patients to predict postoperative dissatisfaction at 2 years. Dissatisfaction was defined as not achieving a minimal clinically important difference in postoperative Knee Society knee and function scores (KSS), Short Form-36 Health Survey [SF-36, divided into a physical component score (PCS) and mental component score (MCS)], and Oxford Knee Score (OKS). Compared to image-only models, both clinical-data only and multimodal models achieved superior performance at predicting dissatisfaction measured by AUC, clinical-data only model: KSS 0.888 (0.866–0.909), SF-PCS 0.836 (0.812–0.860), SF-MCS 0.833 (0.812–0.854), and OKS 0.806 (0.753–0.859); multimodal model: KSS 0.891 (0.870–0.911), SF-PCS 0.832 (0.808–0.857), SF-MCS 0.835 (0.811–0.856), and OKS 0.816 (0.768–0.863). Our findings highlighted that ML models using clinical or multimodal data were capable to predict post-TKA dissatisfaction.

Knee osteoarthritis (OA) is the most common degenerative joint disease which leads to significant disability in the elderly, with a global prevalence of around 23% in individuals aged over 40 years<sup>1</sup>. Total knee arthroplasty (TKA) is recommended for advanced knee OA based on its capacity to reduce pain, improve functional status and quality of life<sup>2</sup>. Despite extensive development and optimization of TKA techniques, it is estimated that 15–20% patients remained dissatisfied after the surgery<sup>3</sup>. Identifying patients who are at risk of dissatisfaction will support better clinical decision making prior to surgery. By proposing alternative management strategies, reduction of outpatient attendances and minimizing unwarranted surgical intervention for this group of patients, we can provide more cost-effective and patient-centered care for patients with knee OA.

The outcome of TKA is commonly assessed objectively with patient-reported outcome measures (PROMs) to measure baseline function preoperatively and patient improvement postoperatively. Several PROMs have been developed and validated worldwide, including Knee Society knee and function scores (KSS)<sup>4</sup>, Short Form-36 Health Survey (SF-36)<sup>5</sup> and Oxford Knee Score (OKS)<sup>6</sup>. These PROMs have corresponding minimal clinically important difference (MCID) thresholds that can be used to determine if a patient achieves perceivable benefits or remains dissatisfied postoperatively.

Machine learning has been applied to predict patient dissatisfaction after TKA. Several studies have used clinical data to predict dissatisfaction by applying tree-based models, achieving an area under the receiver operating characteristic curve (AUC) ranging from 0.60 to 0.95<sup>7</sup>. However, these studies consisted of small sample size, heterogeneous MCID cutoffs and follow-up periods, and used singular functional or mental questionnaires. Mostly notably, no studies have included anterior-posterior (AP) knee radiographs which are used by surgeons to determine disease severity and patient counseling prior to surgery.

The aim of this study was to develop and compare the performance of ML models using image only (AP knee radiographs), clinical data only, and multimodal data to predict postoperative dissatisfaction, defined as not meeting PROM MCID thresholds after TKA at 6-month and 2-year follow-ups, respectively. Specifically, the input included one AP knee radiograph and clinical data per knee collected at preoperative assessment visit of the corresponding patient, and the output was the probability that the patient would not achieve MCID, i.e., dissatisfied with the surgical outcome, measured by various PROMs after TKA at follow-up visits.

<sup>1</sup>Institute of Data Science, National University of Singapore, Singapore, Singapore. <sup>2</sup>School of Computing, National University of Singapore, Singapore, Singapore.

<sup>3</sup>Department of Orthopaedic Surgery, Singapore General Hospital, Singapore, Singapore. <sup>4</sup>These authors contributed equally: Xiaodi Liu, Yingnan Liu.

✉ e-mail: [idsv33@visitor.nus.edu.sg](mailto:idsv33@visitor.nus.edu.sg)

## Results

We included a total of 7224 knee radiographs from 5720 subjects with corresponding clinical data available at baseline and follow-up visits for model development (see Supplementary Figure 1). Details of baseline characteristics are shown in Table 1. The baseline mean age was 67.0 (standard deviation, SD = 7.6) years, 21% (N = 1513) of them were males. The majority of our patients were Chinese (87%, N = 6288), followed by Malay (6.9%, N = 500), Indian (5.1%, N = 370) and other races (0.9%, N = 66). Seventy-nine percent of them were overweight to severely obese, the prevalent comorbidities including hypertension (62%, N = 4501), hyperlipidaemia (49%, N = 3515) and diabetes (18%, N = 1332). The participants had a mean follow-up of 1.9 years (SD = 0.4).

In the internal testing set, the image-only model achieved fair AUCs for the prediction at 2-year follow-up: not achieving MCID in KSS 0.577 [95% confidence interval (95%CI) 0.540–0.612], SF-PCS 0.542 (0.510–0.574), SF-MCS 0.486 (0.455–0.519), and OKS 0.617 (0.557–0.677). Similar model performance was observed in the prediction of 6-month outcomes (see Supplementary Table 1, Supplementary Figure 2).

The clinical-data only model outperformed image-only model for the prediction at 2-year follow-up: not achieving MCID in KSS 0.888 (0.866–0.909), SF-PCS 0.836 (0.812–0.860), not achieving SF-MCS MCID 0.833 (0.812–0.854), and OKS 0.806 (0.753–0.859) (DeLong Test, all  $p < 0.001$ ); as well as at 6-month follow-up: KSS 0.895 (0.873–0.913), SF-PCS 0.835 (0.810–0.859), SF-MCS 0.830 (0.807–0.852), OKS 0.809 (0.763–0.853) (DeLong Test, all  $p < 0.001$ ) (Fig. 1, Table 2).

When concatenating image features with clinical data in the multimodal model, its performance was equivalent but did not reach statistically significance as compared to clinical-data only model at 2-year follow-up: KSS 0.891 (0.870–0.911), SF-PCS 0.832 (0.808–0.857), SF-MCS 0.835 (0.811–0.856), and OKS 0.816 (0.768–0.863) (DeLong test, all  $p > 0.05$ ) (Fig. 1, Table 2). Similar results were observed in the prediction of 6-month outcomes: KSS 0.898 (0.876–0.916), SF-PCS 0.832 (0.808–0.856), SF-MCS 0.835 (0.812–0.857), and OKS 0.802 (0.753–0.848). (see Supplementary Table 1, Supplementary Figure 2) (DeLong Test, all  $p > 0.05$ ).

To interpret the results from image-only model, we selected random cases who underwent bilateral TKA from the testing set (see Supplementary Figure 3). Although models incorporating images can identify pathological changes such as narrowed joint space and osteophytes, it was unable to use these features (presence/severity) to predict patient dissatisfaction. For example, a subject demonstrating similar severity of radiographic features in both knees was satisfied with only one side, while the contralateral side, which presented higher preoperative PROMs, resulted in dissatisfaction after TKA.

Since clinical-data only model achieved equivalent performance with multimodal model, we evaluated the feature importance in the clinical-data only model to determine the most importance contributors. Of the 60 input features, the preoperative score emerged as the most important input variable for predicting postoperative dissatisfaction. The top three features for each PROM are as follows: KSS: preoperative KSS score, preoperative SF-MCS score and height; SF-PCS: preoperative SF-PCS score, age, preoperative SF-35 Q2; SF-MCS: preoperative SF-MCS score, preoperative SF-36 Q6 and preoperative SF-PCS score; OKS: preoperative OKS score; preoperative terminal knee extension and age. The direction of association between each input variables and outcomes at 2-year follow-up is shown in Fig. 2, with similar relationships observed at 6-month follow-up (see Supplementary Figure 4).

## Discussion

To our knowledge, we presented the first and largest interpretable machine learning study comparing three machine learning models that utilized AP knee radiographs only, clinical data only, and multimodal data to predict postoperative dissatisfaction using validated PROMs. Our machine learning models were built, validated, and tested in 5720 subjects at 6-month and 2-year follow-ups, respectively. Among the three models, both clinical-data only and multimodal model obtained excellent performance that

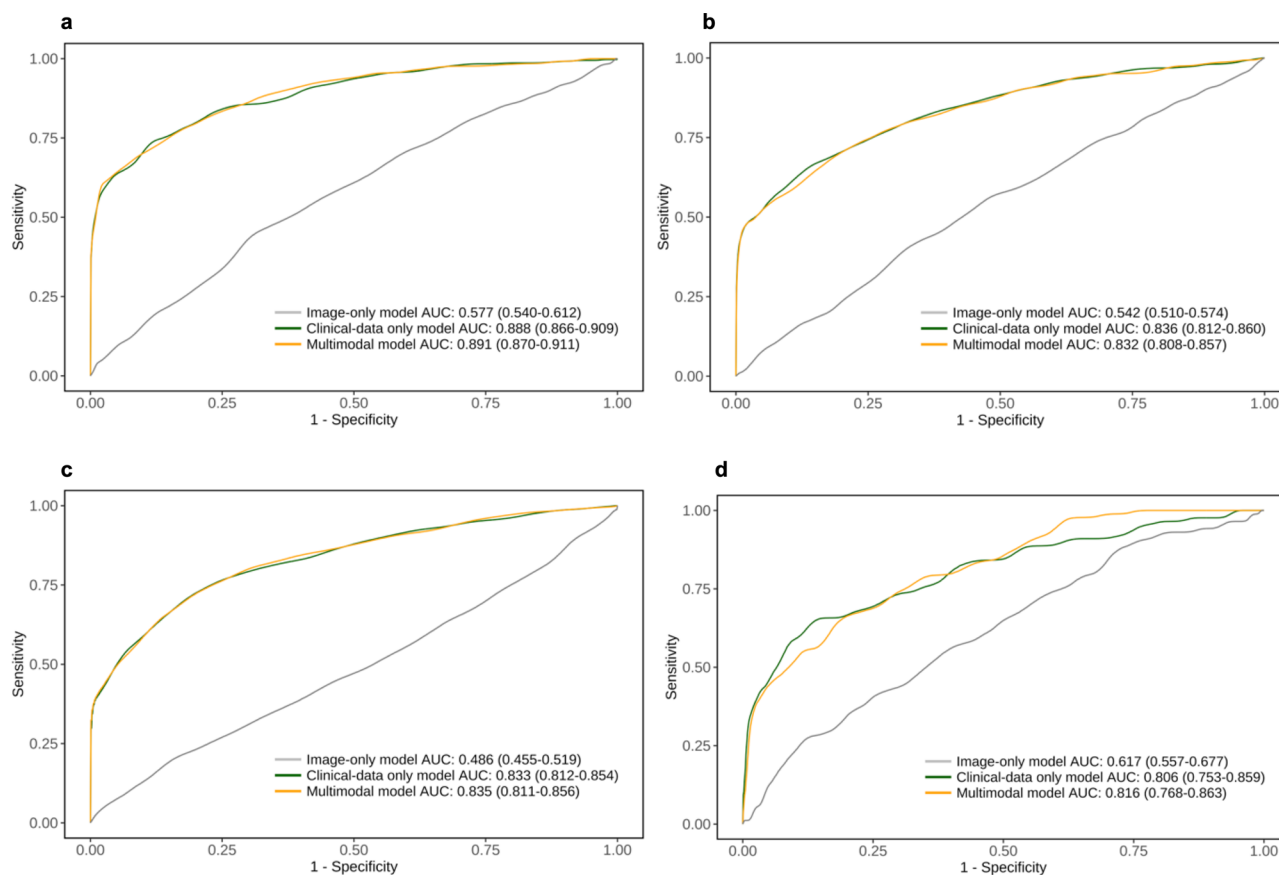
**Table 1 | Baseline characteristics of the patients in training and testing dataset**

| Cohorts                      | Knee dataset | Training set | Testing set |
|------------------------------|--------------|--------------|-------------|
| Number of images             | 7224         | 5778         | 1446        |
| Number of participants       | 5720         | 4575         | 1145        |
| Demographics                 |              |              |             |
| Male                         | 1513 (21)    | 1232 (21)    | 281 (19)    |
| Age, year                    | 67.0 ± 7.6   | 67.2 ± 7.5   | 67.0 ± 7.7  |
| BMI                          |              |              |             |
| Normal (BMI < 24)            | 1462 (21)    | 1151 (20)    | 311 (22)    |
| Overweight (24 ≤ BMI < 28)   | 2693 (37)    | 2138 (37)    | 555 (38)    |
| Obese (28 ≤ BMI < 32)        | 1954 (27)    | 1609 (28)    | 345 (24)    |
| Severely obese (BMI ≥ 32)    | 1115 (15)    | 880 (15)     | 235 (16)    |
| Ethnicity                    |              |              |             |
| Chinese                      | 6288 (87)    | 5024 (87)    | 1264 (87)   |
| Malay                        | 500 (6.9)    | 410 (7.1)    | 90 (6.2)    |
| Indian                       | 370 (5.1)    | 295 (5.1)    | 75 (5.2)    |
| Others                       | 66 (0.9)     | 49 (0.8)     | 17 (1.2)    |
| Clinical characteristics     |              |              |             |
| Hypertension                 | 4501 (62)    | 3,619 (63)   | 882 (61)    |
| Diabetes                     | 1332 (18)    | 1094 (19)    | 238 (16)    |
| Hyperlipidemia               | 3515 (49)    | 2828 (49)    | 687 (48)    |
| Ischemic heart disease       | 596 (8.3)    | 483 (8.4)    | 113 (7.8)   |
| Arthritis other than knee OA | 240 (3.3)    | 188 (3.3)    | 52 (3.6)    |
| Depression                   | 66 (0.9)     | 43 (0.7)     | 23 (1.6)    |
| Outcomes at 6 months         |              |              |             |
| Not achieve KSS MCID         | 2035 (28)    | 1631 (28)    | 404 (28)    |
| Not achieve OKS MCID         | 857 (12)     | 556 (9.6)    | 134 (9.3)   |
| Not achieve SF-MCS MCID      | 5309 (73)    | 4262 (74)    | 1047 (72)   |
| Not achieve SF-PCS MCID      | 2734 (38)    | 2196 (38)    | 538 (37)    |
| Outcomes at 2 years          |              |              |             |
| Not achieve KSS MCID         | 1808 (25)    | 1432 (25)    | 376 (26)    |
| Not achieve SF-MCS MCID      | 5172 (72)    | 4137 (72)    | 1035 (72)   |
| Not achieve SF-PCS MCID      | 2320 (32)    | 1841 (32)    | 479 (33)    |
| Not achieve OKS MCID         | 559 (7.7)    | 346 (6.0)    | 88 (6.1)    |

Data are mean (SD) or n (%).

BMI body mass index, KSS Knee Society knee and function scores, MCID minimal clinical important difference, MCS mental component score, OA osteoarthritis, OKS Oxford Knee Score, PCS physical component score, SD, standard deviation, SF Short Form-36 Health Survey.

outperformed the image-only model, measured by AUC and its 95% CI, indicating that incorporation of AP knee radiographs commonly used by surgeons to monitor disease progression and preoperative TKA candidacy counseling could only add limited values on predicting postoperative dissatisfaction. We also identified that preoperative function scores are the most important clinical features that correlate with postoperative dissatisfaction.



**Fig. 1 | Receiver operating characteristic curves for the internal testing set on predicting patient dissatisfaction at 2-year follow-up.** ROC curves of image-data only model, clinical-data only model and multimodal model for predicting each PROM MCID: **a** Not achieve KSS MCID, **b** Not achieve SF-PCS MCID, **c** Not

achieve SF-MCS MCID. **d** Not achieve OKS MCID. KSS Knee Society knee and function scores, MCS mental component score, OKS Oxford Knee Score, PCS physical component score, SF Short Form-36 Health Survey.

**Table 2 | Model performance for the internal testing set on predicting patient dissatisfaction at 2-year follow-up**

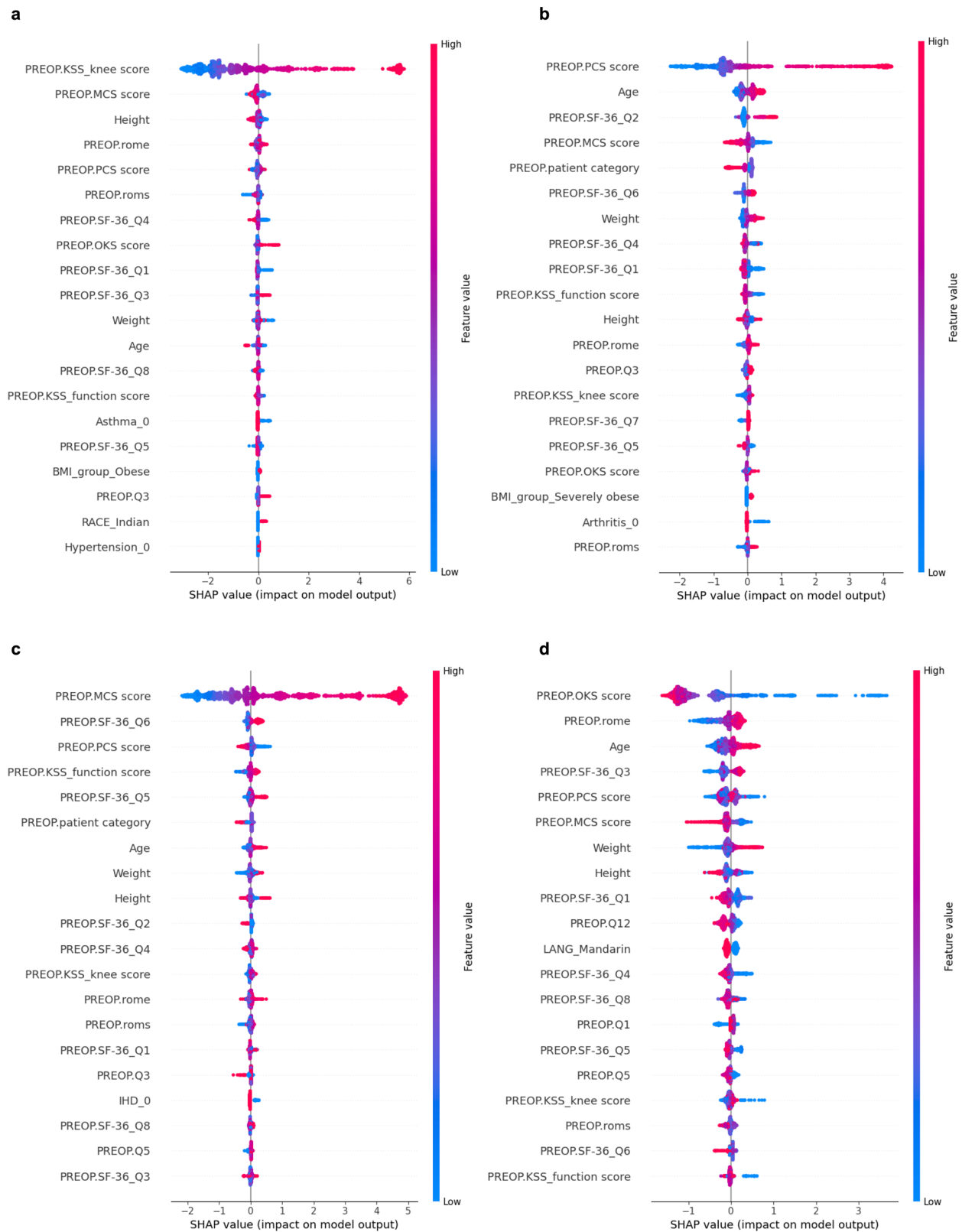
| Outcomes                   | Models             | AUC                 | F1 score            | Precision           | Recall              |
|----------------------------|--------------------|---------------------|---------------------|---------------------|---------------------|
| Not achieve KSS MCID       | Image-only         | 0.577 (0.540–0.612) | 0.393 (0.355–0.432) | 0.311 (0.276–0.346) | 0.535 (0.483–0.588) |
|                            | Clinical-data only | 0.888 (0.866–0.909) | 0.713 (0.676–0.748) | 0.685 (0.640–0.730) | 0.744 (0.700–0.790) |
|                            | Multimodal         | 0.891 (0.870–0.911) | 0.697 (0.658–0.730) | 0.674 (0.627–0.717) | 0.721 (0.672–0.769) |
| Not achieve SF-36 PCS MCID | Image-only         | 0.542 (0.510–0.574) | 0.333 (0.290–0.373) | 0.366 (0.315–0.414) | 0.306 (0.263–0.350) |
|                            | Clinical-data only | 0.836 (0.812–0.860) | 0.676 (0.638–0.711) | 0.717 (0.672–0.761) | 0.639 (0.593–0.684) |
|                            | Multimodal         | 0.832 (0.808–0.857) | 0.659 (0.623–0.696) | 0.694 (0.648–0.737) | 0.627 (0.583–0.676) |
| Not achieve SF-36 MCS MCID | Image-only         | 0.486 (0.455–0.519) | 0 (0–0)             | 0 (0–0)             | 0 (0–0)             |
|                            | Clinical-data only | 0.833 (0.812–0.854) | 0.788 (0.765–0.810) | 0.910 (0.888–0.932) | 0.695 (0.665–0.724) |
|                            | Multimodal         | 0.835 (0.811–0.856) | 0.770 (0.746–0.793) | 0.912 (0.889–0.932) | 0.667 (0.636–0.698) |
| Not achieve OKS MCID       | Image-only         | 0.617 (0.557–0.677) | 0.138 (0.101–0.177) | 0.081 (0.058–0.107) | 0.441 (0.349–0.544) |
|                            | Clinical-data only | 0.806 (0.753–0.859) | 0.367 (0.287–0.434) | 0.267 (0.201–0.332) | 0.590 (0.479–0.685) |
|                            | Multimodal         | 0.816 (0.768–0.863) | 0.424 (0.329–0.511) | 0.459 (0.349–0.573) | 0.397 (0.293–0.494) |

Data are mean (95% CI).

AUC area under the receiver operating characteristic curve, CI confidence interval, KSS Knee Society knee and function scores, MCID minimal clinical important difference, MCS mental component score, OKS Oxford Knee Score, PCS physical component score, SF Short Form-36 Health Survey.

A recent review has summarized the existing evidence and developments of machine learning models in the prediction of postoperative satisfaction<sup>7</sup>, showing that weakness of previous studies included inconsistent performance (AUC 0.60–0.95), limited sample size (around 400–6400), various clinical outcomes, used only clinical data as the input

data, and heterogenous follow-up period. For instance, one cross-sectional NHS-based study involving 34,100 observations reported J-statistic (sensitivity + specificity – 1) of 0.31 for predicting postoperative OKS score and 0.57 for quality of life<sup>8</sup>, the other study following 6480 patients who received TKA at 2-year visit demonstrated an AUC of 0.6–0.89<sup>9</sup>. Our findings have



**Fig. 2 | Feature importance plots on predicting patient dissatisfaction at 2-year follow-up.** SHAP explanation of clinical-data only model for predicting each PROM MCID: **a** Not achieve KSS MCID. **b** Not achieve SF-PCS MCID. **c** Not achieve SF-MCS MCID. **d** Not achieve OKS MCID. KSS Knee Society knee and function scores,

MCID minimal clinically important difference, MCS mental component score, OKS Oxford Knee Score, PCS physical component score, PREOP preoperative, rome terminal flexion, roms terminal extension.

addressed the issues of inconsistent model performance, limited sample size, and heterogeneous follow-up periods, offering valuable and interpretable evidence by incorporating AP knee radiograph for identifying patients who are unlikely to benefit from TKA, therefore provide precise management in the clinical settings.

Over 700,000 TKA are performed every year in the United States<sup>10</sup>, with around 15–20% of patients would be dissatisfied postoperatively for unknown reasons. The recommendation for TKA relies fundamentally on medical history of chronic mechanical knee pain which limits function, physical examination suggesting internal derangement of the knee and severe pathological changes on preoperative knee radiographs. Radiographs remain the most widely used tool to assess the severity of structural damage such as joint narrowing space and osteophytes. However, the relationship between radiographic findings with PROMs are inconsistent in previous studies. One study identified that increased Kellgren–Lawrence (KL) score was associated with worsen general health, physical function and pain<sup>11</sup>, while other two studies found weak or no relationship between KL grading with PROMs<sup>12,13</sup>, indicating the severity of radiographic features cannot fully reflect patient's functional, mental or quality of life status. Our study, from a machine learning standpoint, shows that although the model can identify pathological patterns such as narrowed joint space in the AP knee images, it inadequately associates with the clinical outcomes as measured by PROMs after TKA.

Although standard knee radiographs include AP view, lateral view, and the Merchant view, most of the orthopedic machine learning studies to date mainly used the AP view for joint segmentation, early detection and severity classification, with only two studies exploring the feasibility of using the lateral view for early detection of OA knee<sup>14,15</sup>. Emerging evidence has shown the patellofemoral joint space narrowing, assessed by a lateral or Merchant view, is associated with poorer PROMs<sup>16</sup>. In concordance with previous findings, our study suggests that although the machine learning models were capable to learn certain pathological patterns from the AP knee radiograph, the image alone only offered limited information on predicting patient dissatisfaction after TKA, highlighting that surgeons should not solely use AP knee radiographs when indicating patients for TKA. One of the possible explanations is that the two-dimensional bony knee structure in the AP view, without information on the actual degree of cartilage loss, is incapable of reflecting preoperative function<sup>17</sup> that is highly associated with postoperative dissatisfaction. Future studies incorporating the lateral or the Merchant view are warranted to improve the accuracy of model prediction.

The postoperative dissatisfaction is a multifactorial outcome that correlates with patient expectation, medical history, female sex, ethnicity, and functional status<sup>18</sup>. Recent studies have raised concerns that patients' mental health before surgery may be negatively associated with the postoperative dissatisfaction. In a longitudinal prospective study following over 1200 patients who received TKA over a mean of 5 years, patients with a higher preoperative mental health score presented better physical functions scores at pre- and post-operative visits<sup>19</sup>. In another pilot randomized controlled trial, preoperative psychological intervention improved patient satisfaction at 2-year follow-up after TKA among patients with depression<sup>20</sup>. Our results showed that preoperative mental health status was ranked as a top 4 feature in predicting dissatisfaction measured by KSS, SF-PCS, and SF-MCS, indicating psychological distress may play an important role on patient dissatisfaction. Assessment of both functional and mental health status coupled with timely intervention may be beneficial in the clinical management of OA knee.

Several clinical rating systems have been widely used to describe the outcome of TKA, among which KSS is the most utilized scoring system<sup>21</sup>. The KSS objectively assesses functional status, whereas SF-36 measures quality of life, and OKS evaluates patient's opinion on their outcome that is more correlated with pain but less with the functional status<sup>22</sup>. In addition, the concepts of measuring satisfaction, expectation fulfillment, patient acceptable symptoms state (PASS) and MCID are not equivalent. Unfortunately, none of these PROMs can comprehensively address the changes in pain, physical, function, and mental status after TKA. Future studies should

be directed toward designing a comprehensive scoring system in order to understand patient-oriented outcomes.

Our study has identified top performing features, namely preoperative PROM scores, range of motion, and age from validated PROMs that can predict patient dissatisfaction. These clinical parameters could form the basis of machine learning models used in either primary care or specialist care to select or indicate patients for TKA. These novel models, instead of a single-view AP knee radiograph can be used to predict patient dissatisfaction and counsel patients on the probability of a favorable postoperative outcome. Through careful patient selection, this will lead to improved overall outcomes for TKA patients and enhance cost-effectiveness of the procedure.

Our study has some limitations. We performed internal validation within one institution, further external validation with multi-ethnicity cohorts is desirable. Second, our study may have selection bias since we excluded patients with surgical history. Third, our model is a classification one that cannot estimate the exact postoperative PROM scores. Fourth, the models did not take into account postoperative radiographic assessments where poorly positioned components may affect outcomes.

In summary, we developed and evaluated three machine learning models utilizing image-data only, clinical-data only, and multimodal data to predict patient dissatisfaction at 6-month and 2-year follow-ups. This machine learning study is the first to utilize and compare whether incorporating AP knee radiograph can predict patient dissatisfaction after TKA. We showed that both clinical-data only and multimodal model presented excellent performance on predicting postoperative dissatisfaction. Our model could serve as a valuable evaluation tool to differentiate patients with knee OA who will not benefit from TKA in order to achieve the precision medicine in the clinical settings. Future work to incorporate multi-view knee radiographs to enhance the model performance is warranted.

## Methods

### Study design and participants

In this longitudinal, retrospective, single-center study, we trained, validated, and tested machine learning models using AP knee radiographs only, clinical data only, and multimodal data for predicting postoperative dissatisfaction from a retrospective joint replacement registry in Singapore General Hospital, Singapore. Patients who underwent either TKA or minimally-invasive TKA (MIS-TKA) for diagnoses of knee OA from 2006 to 2016 were included. Exclusion criteria were: (1) prior surgery of the knee; (2) underwent non-TKA procedure; (3) incomplete clinical data.

We collected anonymized preoperative clinical assessments, PROMs, and weight-bearing AP knee radiographs at baseline, with PROMs repeated at 6 months and 2 years postoperatively. The ethics approval was obtained from the SingHealth centralized institutional review board (Reference number: CIRB 2019/2878), informed consent was waived by the ethics committee since all the data were extracted anonymously from routine clinical practice. We followed the TRIPOD guideline to report our study.

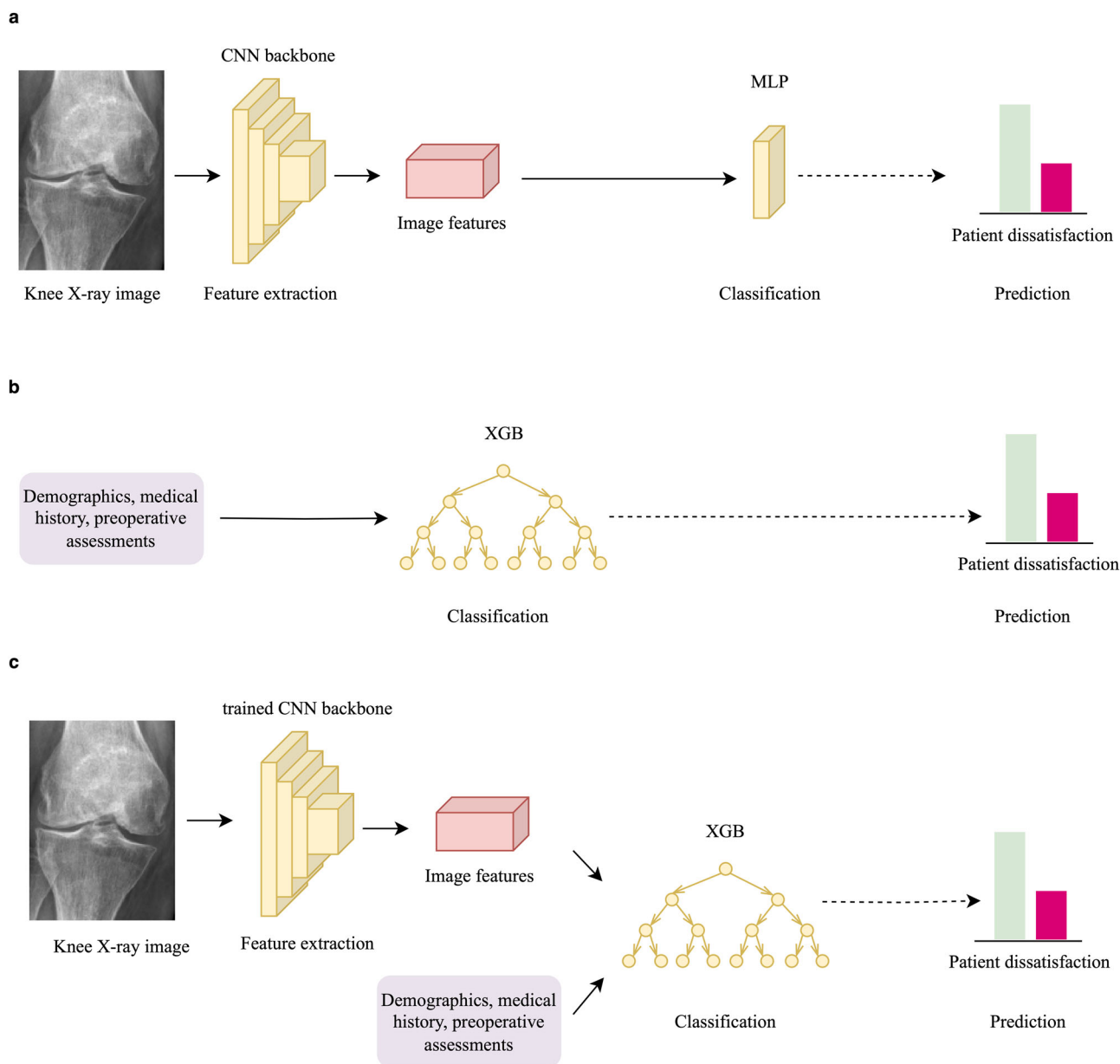
### Data preparation

For clinical data, all continuous variables were normalized, and categorical variables were one-hot encoded.

For each AP knee radiograph in digital imaging and communications in medicine (DICOM) format, we extracted a region of interest of the knee joint using BoneFinder<sup>®</sup> tool (<http://bone-finder.com/>)<sup>23</sup>. Subsequently, we used contrast-limited adaptive histogram equalization<sup>24</sup> to enhance the image contrast. We horizontally flipped all right knee radiographs to have a similar view with the left knee radiographs and rescaled all the images to 224 × 224 pixels.

We performed the dataset splitting to ensure no samples overlapping at the patient level in training and testing sets (80:20). We used the balanced sampling to handle class imbalance between dissatisfied *versus* satisfied patients.





**Fig. 3 | Overview of the proposed machine learning models.** **a** The image-only model feeds with AP knee radiographs. **b** The clinical-data only model feeds with clinical data. **c** The multimodal model incorporating the information from both AP knee radiographs and clinical data. CNN convolutional neural network, MLP multilayer perceptron, XGB Extreme Gradient Boosting.

**Outcomes**

The main goal of this study was to develop and compare the performance of machine learning models using image data only, clinical data only, and multimodal data to predict postoperative dissatisfaction. Several PROMs were recorded by the physiotherapists at preoperatively, at 6 months and at 2 years postoperatively: KSS<sup>4</sup> [ranging from 0 (extreme knee problem) to 100 (no knee problem)], SF-36<sup>5</sup>, and OKS<sup>6</sup> [ranging from 0 (most severe symptoms) to 48 (least severe symptoms)]. The SF-36 score was aggregated into a physical component summary (PCS) [ranging from 0 (worst health status) to 100 (best health status)] and a mental component summary (MCS) [ranging from 0 (worst health status) to 100 (best health status)].

We calculated the difference of each PROM between follow-up visits and preoperative assessment, and dichotomized patients into two groups: dissatisfied (did not achieve MCID) *versus* satisfied (achieved MCID) according to each PROM MCID. We used published MCID values as the cutoff: KSS (34.5 points)<sup>25</sup>, SF-36 (PCS: 10 points and MCS: 10 points)<sup>26</sup>, and

OKS (5 points)<sup>27</sup>. The primary outcomes were patient dissatisfaction quantified by each MCID at 2-year follow-up, and the secondary outcomes were patient dissatisfaction at 6-month follow-up.

We used the area under the received operating characteristic curve (AUC) as the main model-discriminative performance metric. We also report other clinically relevant metrics: F1 score, precision, and recall.

**Model development and test**

We developed three machine learning models to predict patient dissatisfaction after TKA: (a) a model that used knee radiograph only (Fig. 3a), (b) a model that utilized clinical data only (Fig. 3b), (c) a model that integrated both knee radiograph and clinical data for its predictions (Fig. 3c).

For the image-only model, we used a convolution neural network (CNN) to extract features for classification. The ConvNeXt-Tiny<sup>28</sup> (the best model among ResNet101, ResNeXt, ConvNeXt, and Vision Transformer

based on the training set) pretrained with the ImageNet dataset (<http://www.image-net.org/>) was used as the CNN backbone.

We trained clinical-data only model by Extreme Gradient Boosting (XGB) algorithm<sup>29</sup> [the best model among XGB, Random Forest, Support Vector Machine, decision tree and multilayer perceptron (MLP) based on the training set], which comprised demographics, medical history, pre-operative assessments and preoperative PROM scores (see Supplementary Table 2). We applied 5-fold internal cross validation (4 folds for training and 1 fold for validation, in turn) in the training set for hyperparameter tuning. Once the optimal hyperparameter values were selected, the actual XGB model was trained on the whole training set. The hyperparameters were optimized for the number of estimators, maximum depth, learning rate, and subsample strategies (see Supplementary Table 3). We selected the best hyperparameter and epoch achieved the highest average performance of AUC on the validation fold for final training, then evaluated on the testing set.

For the multimodal model, after the same CNN architecture was trained to extract image features, we concatenated the image features with clinical features and used them to train a XGB classifier via a joint training approach. Specifically, the CNN was jointly trained with MLP-extracted clinical features to extract complementary image features for further prediction. We used Adam optimizer to update each CNN architecture. In image-only and multimodal models, we used the same 5-fold cross validation as previously described within the training set for hyperparameter tuning in CNN (learning rate, weight decay, batch size, and best epoch) and XGB (same hyperparameters searched in clinical-data only model), respectively. The model was trained for 10 epochs. We used the cross-entropy loss function, learning rate of  $5 \times 10^{-5}$ , batch size of 32, weight decay of 0, and a balanced class sample for training.

Our models were implemented using PyTorch and xgboost. We conducted all the training on a machine equipped with a NVIDIA A100 SXM4 Tensor Core graphics processing unit and 80 GB available video random-access memory.

### Model interpretation

We used SHAP<sup>30</sup> to identify the key clinical features and gradient-weighted class activation mapping (Grad-CAM)<sup>31</sup> to visualize where the model is focusing when making predictions on patient dissatisfaction.

### Statistical analysis

To estimate the confidence interval (CI) for each performance metric, we used n-out-of-n bootstrap with replacement, repeating it for 1000 times to calculate the 95% CI. We estimated 95% CI using the 2.5 and 97.5 percentiles of the empirical distribution of the corresponding metric. We compared the model performance in three models: (1) image-only, (2) clinical-data only, (3) multimodal using the DeLong's method as well as the overlap of 95% CI of mean AUC scores in each model. *P* values < 0.05 were considered as statistically significance.

Data analysis and modeling were performed using Python 3.9.17 (Python Software Foundation, Wilmington, DE, USA), the Anaconda Distribution (Anaconda, Inc., Austin, TX, USA), and R software 4.2.2 (R Foundation for Statistical Computing, Vienna, Austria, 2022).

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The dataset used in this study are protected patient information that only restricted access is allowed.

### Code availability

The source code for training the machine learning models used in this study is available at: <https://github.com/NancyQuris/OAKneeML>.

Received: 21 February 2024; Accepted: 17 September 2024;

Published online: 30 September 2024

## References

- Cui, A. et al. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *EClinicalMedicine* **29**, 100587 (2020).
- Panel, N. C. NIH Consensus Statement on total knee replacement December 8-10, 2003. *J. Bone Jt. Surg. Am.* **86**, 1328–1335 (2004).
- DeFrance, M. J. & Scuderi, G. R. Are 20% of patients actually dissatisfied following total knee arthroplasty? A systematic review of the literature. *J. Arthroplast.* **38**, 594–599 (2023).
- Noble, P. C. et al. Development of a new knee society scoring system. *Clin. Orthop. Relat. Res.* **470**, 20–32 (2012).
- Ware, J. E. & Sherbourne, C. D. The MOS 36-Item short-form health survey (SF-36). *Méd. Care* **30**, 473–483 (1992).
- Dawson, J., Fitzpatrick, R., Murray, D. & Carr, A. Questionnaire on the perceptions of patients about total knee replacement. *J. Bone Jt. Surg. Br.* **80-B**, 63–69 (1998). *Vol.*
- Lee, L. S. et al. Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: a review. *Arthroplasty* **4**, 16 (2022).
- Huber, M., Kurz, C. & Leidl, R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Méd. Inform. Decis. Mak.* **19**, 3 (2019).
- Fontana, M. A., Lyman, S., Sarker, G. K., Padgett, D. E. & MacLean, C. H. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clin. Orthop. Relat. Res.* **477**, 1267–1279 (2019).
- Katz, J. N., Arant, K. R. & Loeser, R. F. Diagnosis and treatment of hip and knee osteoarthritis. *JAMA* **325**, 568–578 (2021).
- Pereira, D. et al. Knee and hip radiographic osteoarthritis features: differences on pain, function and quality of life. *Clin. Rheumatol.* **35**, 1555–1564 (2016).
- Özden, F. et al. The relationship of radiographic findings with pain, function, and quality of life in patients with knee osteoarthritis. *J. Clin. Orthop. Trauma* **11**, S512–S517 (2020).
- Innmann, M. M. et al. What is the correlation between clinical and radiographic findings in patients with advanced osteoarthritis of the knee? *J. Clin. Med.* **12**, 5420 (2023).
- Bayramoglu, N., Nieminen, M. T. & Saarakkala, S. Automated detection of patellofemoral osteoarthritis from knee lateral view radiographs using deep learning: data from the Multicenter Osteoarthritis Study (MOST). *Osteoarthr. Cartil.* **29**, 1432–1447 (2021).
- Bayramoglu, N., Nieminen, M. T. & Saarakkala, S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis. *Int. J. Méd. Inform.* **157**, 104627 (2022).
- Macri, E. M. et al. Relation of patellofemoral joint alignment, morphology, and radiographic osteoarthritis to frequent anterior knee pain: data from the multicenter osteoarthritis study. *Arthritis Care Res.* **72**, 1066–1073 (2020).
- Rooij, Mde. et al. Prognosis of pain and physical functioning in patients with knee osteoarthritis: a systematic review and meta-analysis. *Arthritis Care Res.* **68**, 481–492 (2016).
- Rodriguez-Merchan, E. C. Patient satisfaction following primary total knee arthroplasty: contributing factors. *Arch. Bone Jt. Surg.* **9**, 379–386 (2020).
- Lavernia, C. J., Villa, J. M. & Iacobelli, D. A. What is the role of mental health in primary total knee arthroplasty? *Clin. Orthop. Relat. Res.* **473**, 159–163 (2015).
- Geng, X. et al. A randomized controlled trial of psychological intervention to improve satisfaction for patients with depression undergoing TKA: a 2-year follow-up. *J. Bone Jt. Surg.* **103**, 567–574 (2020).

21. Theodoulou, A., Bramwell, D. C., Spiteri, A. C., Kim, S. W. & Krishnan, J. The use of scoring systems in knee arthroplasty: a systematic review of the literature. *J. Arthroplast.* **31**, 2364–2370.e8 (2016).
22. Hove, R. P., van, Brohet, R. M., Royen, B. Jvan & Nolte, P. A. High correlation of the Oxford Knee Score with postoperative pain, but not with performance-based functioning. *Knee Surg. Sports Traumatol. Arthrosc.* **24**, 3369–3375 (2016).
23. Lindner, C. et al. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE Trans. Méd. Imaging* **32**, 1462–1472 (2013).
24. Pisano, E. D. et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J. Digit. Imaging* **11**, 193 (1998).
25. Jacobs, C. A. & Christensen, C. P. Correlations between knee society function scores and functional force measures. *Clin. Orthop. Relat. Res.* **467**, 2414–2419 (2009).
26. Escobar, A. et al. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. *Osteoarthritis Cartil.* **15**, 273–280 (2007).
27. Clement, N. D., MacDonald, D. & Simpson, A. H. R. W. The minimal clinically important difference in the Oxford knee score and Short Form 12 score after total knee arthroplasty. *Knee Surg., Sports Traumatol., Arthrosc.* **22**, 1933–1939 (2014).
28. Liu, Z. et al. A ConvNet for the 2020s. In *Proc. IEEE/CVF Conference Computer Vision Pattern Recognition (CVPR)* 11966–11976 (2022).
29. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. *arXiv*. <https://doi.org/10.1145/2939672.2939785> (2016).
30. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv*. <https://doi.org/10.48550/arxiv.1705.07874> (2017).
31. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision (ICCV)* 618–626. <https://doi.org/10.1109/iccv.2017.74> (2017).

## Acknowledgements

This study was funded by the Singhealth Duke-NUS Musculoskeletal Sciences Academic Clinical Programme Nurturing Clinician Scientist Scheme (NCSS) Research Support Grant. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. We thank Max Jun Neng Tan, Jiale Hu and Evelyn Welter Lim for image pre-processing.

## Author contributions

X.D.L. designed and implemented the methods, reviewed the literature and wrote the article. Y.N.L. designed and implemented the methods and interpreted the data. M.L.L. and W.H. analyzed and interpreted the data. M.H.L.L. collected and interpreted the data and revised the article. All authors reviewed and approved the manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-024-01265-8>.

**Correspondence** and requests for materials should be addressed to Xiaodi Liu.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024