DIGITAL HEALTH

Automated machine learning-based prediction of the progression of knee pain, functional decline, and incidence of knee osteoarthritis in individuals at high risk of knee osteoarthritis: Data from the osteoarthritis initiative study

Tianrong Chen and Calvin Kalun Or [iD]

Abstract

Objective: This study aimed to examine the performance of machine learning models in predicting the progression of knee pain, functional decline, and incidence of knee osteoarthritis (OA) in high-risk individuals, with automated machine learning (AutoML) being used to automate the prediction process.

Design: There were four stages in the process of our AutoML-integrated prediction. Stage 1−Data preparation: The data of 3200 eligible individuals in the Osteoarthritis Initiative (OAI) study who were considered at high risk of knee OA at the baseline visit were extracted and used. Specifically, 1094 variables from the OAI study were used to predict the changes in knee pain, physical function, and incidence of knee OA (i.e. the first occurrence of frequent knee symptoms and definite tibial osteophytes (Kellgren and Lawrence grade $\geq$2)) over a 9-year period. Stage 2−Model training: The AutoML approach was used to automatically train nine widely used machine learning (ML) models. Stage 3−Model testing: The AutoML approach was used to automatically test the performance of the ML models. Stage 4−Selection of important input variables: The AutoML approach automated the process of computing the importance scores of all input variables and identifying the most important ones, using the technique of permutation feature importance.

Results: Using the AutoML approach, the weighted ensemble model and the CatBoost model showed the best performance among all nine ML models. For the prediction of each outcome in each year, the five most important input variables were identified, most of which were obtained from self-reported questionnaire surveys and radiographic imaging reports.

Conclusion: The AutoML approach has shown potential in automating the process of using ML models to predict long-term changes in knee OA-related outcomes. Its use could support the deployment of ML solutions, facilitating the provision of personalized interventions to prevent the deterioration of knee health and incident knee OA.

Keywords

Knee osteoarthritis, knee pain, physical function, prediction, automated machine learning

Introduction

Knee osteoarthritis (OA) is a leading cause of knee pain, immobility, and poor quality of life, which affects more than 20% of middle-aged and older adults worldwide.[1–5] Many more people will be at risk of knee OA in future because of aging and obesity,[6] so it is critical for OA to

Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong, China

Corresponding author:
Calvin Kalun Or, Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Room 8-7, 8/F, Haking Wong Building, Hong Kong, China.
Email: klor@hku.hk

be prevented or to be detected early. Typically, diagnosis of knee OA relies on patient-reported symptoms and evidence from X-radiographic images, but individuals may not notice subtle changes in their knee health, and X-radiography has limitations in detecting early disease.[7–9] As a result, clinicians may fail to ensure that their patients take necessary precautions to prevent the progression of knee pain, functional decline, and the occurrence of knee OA.

Artificial intelligence-based prognostic models can be developed and may enable the prediction of incident knee OA, for which, machine learning (ML) algorithms such as logistic regression[10] and k-nearest neighbors (k-NN)[11] have been used. However, there are limitations to such approaches. First, the deployment of high-performance models demands manual preprocessing of data, feature engineering, model selection, hyperparameter tuning, and model testing, all of which can be complex and time-consuming.[12] As such, clinicians who lack knowledge and experience in programming may struggle to use predictive models. Second, the risk factors that were used as the input variables in previously developed predictive models may not be the most important variables to use for accurate prediction.[13] Identifying the most important factors for precise prediction is an ongoing challenge in this field.

Automated machine learning (AutoML) can be used to automatically execute the processes of implementing ML models to generate predictions (e.g. selecting and combining ML models, optimizing hyperparameter settings, and achieving optimal performance) in a more flexible, robust, and efficient manner than by using traditional ML techniques.[14–16] With the assistance of AutoML, clinicians would be able to devise and deploy ML solutions without the need to expend extensive time and effort in model training, testing, and identification of important input variables.[14] Despite these benefits, little research has been conducted on the use of AutoML to assist in the process of implementing ML models to predict knee health conditions in individuals at high risk of knee OA.

Therefore, the aim of the present study was to apply an AutoML approach to automatically manage the processes of implementing ML models and to examine the performance of these models in predicting the progression of knee pain, functional decline, and incidence of knee OA in individuals at high risk of knee OA. Data from the Osteoarthritis Initiative (OAI) study, which comprise over 3000 eligible participants and over 1000 input variables (i.e. clinical and imaging data collected from the OAI clinical visits), were used to train and test the ML models, and the most important variables for the predictions were identified.

Methods

Study sample

The OAI (https://nda.nih.gov/oai) is a longitudinal study of the natural progression of knee OA in 4796 individuals aged 45 to 79 years from four clinical centers in the United States. These individuals are examined annually, and the OAI clinical data collected at the baseline visit and nine follow-up visits have been made public. Ethical approval was obtained from the Institutional Review Boards of four OAI clinical sites, located at Baltimore, Maryland; Columbus, Ohio; Pittsburgh, Pennsylvania; and Pawtucket, Rhode Island. All participants included in the OAI study provided written informed consent.

In the current study, we used the OAI study's eligibility criteria to include individuals at high risk of knee OA. The OAI study defines individuals at high risk of knee OA (i.e. the incidence cohort) as those who had risk factors at baseline but were not diagnosed with symptomatic tibial-femoral OA, that is, those who did not have frequent knee symptoms (pain, soreness, or stiffness in or around the knee on most days for at least 1 month in the past 12 months) and definite tibial osteophytes (Kellgren and Lawrence imaging grade ≥ 2) in the same knee.[17] The risk factors used by the OAI study to identify the individuals at high risk of knee OA ($n = 3284$) from all of the participants include age above 70, overweight, presence of knee symptoms during the past 12 months, previous knee injury, previous knee surgery, family history of total knee replacement for knee OA, and Heberden's node in the hand.[17]

Knee OA-related outcomes

Three knee OA-related outcomes were used in our predictions: progression of knee pain, functional decline, and incidence of knee OA. We used the participants' scores on the pain subscale and the physical function subscale of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)[18] to predict the progression of knee pain and functional decline over nine years. The pain subscale contains five items on knee pain: knee pain during walking, while using stairs, while in bed, while sitting or lying, and while standing upright. The physical function subscale of the WOMAC comprises 17 items measuring the difficulty of performing daily activities, such as using stairs, rising from sitting, bending, shopping, and performing domestic duties. The items are scored on a 5-point Likert scale ranging from 0 (none) to 4 (extreme), with a maximum score of 20 for the pain subscale and 68 for the physical function subscale. A high score indicates worse knee health. With respect to the prediction of incident knee OA, the participants included in this study had no knee OA at baseline, and we predicted whether they exhibited knee OA at the follow-up visits each year. According to the OAI study, the incidence of knee OA is defined as the first occurrence of frequent knee symptoms (pain, soreness, or stiffness in or around the knee on most days for at least 1 month in the past 12 months) and definite tibial osteophytes (Kellgren and Lawrence imaging grade ≥ 2) in the same knee.[17]

Input variables used in our predictions

We first excluded the administrative variables from the OAI study data, such as data collection date and staff identifications, and finally extracted 1094 variables, which we subsequently used in our prediction of the progression of knee pain, functional decline, and incidence of knee OA in the participants over nine years. The input variables we used to generate our predictions comprised demographic information, joint symptoms and functions, quality of life, medical history, surgery history, performance of physical examinations, physical activity, dietary nutrition intake, knee examination, and assessments of knee X-ray images.

AutoML-based prediction

For each of the three outcomes, we used the baseline data of both knees as input variables and predicted the conditions of the right and left knees separately for each follow-up visit. This approach would enable individuals to have a better understanding of their knee health, such as which knee may experience deterioration in specific years in the future, allowing for the provision of more personalized interventions. Figure 1 presents the four stages in the ML pipeline of AutoML applied in our prediction task. We applied AutoGluon[15] (an open source AutoML platform) to execute the processes of model training and testing in an automatic manner, which was shown to be faster, more robust, and much more accurate than many other AutoML platforms.[15] The details of the four stages are described below.

Stage 1—Data preparation. The data records of the OAI's participants who were considered at high risk of knee OA at the baseline visit were located, and input variables and their outcome data were extracted. The outcome data of individuals that were acquired after knee replacement surgery or the outcome data of those who had died were not extracted and used. After excluding the participants with no outcome data, the remaining sample was used for the training and testing of the predictive models. Their data were randomly divided into a training set (90% of the data) to fit the parameters of the predictive models, and a testing set (10% of the data) to examine the performance of the models by comparing the predicted values with the true values of the outcomes. The procedure of data splitting was performed 10 times, along with the following processes of model training and testing, to alleviate the problem of overfitting or selection bias.[19] The mean value of the performances of each model across these 10 runs were reported.

Stage 2—Model training. We applied a random forest algorithm, an extra-trees algorithm, a gradient-boosting algorithm (using five implementations), a *k*-NN algorithm, and a weighted ensemble approach[15] to develop regression models to predict the progression of knee pain and functional decline, and applied classification models to predict the incidence of knee OA. We used AutoML to automatically search for hyperparameters to optimize the performance of each model in the training set.

Stage 3—Model testing. We calculated the root-mean-square errors (RMSEs) to evaluate the performance of the models
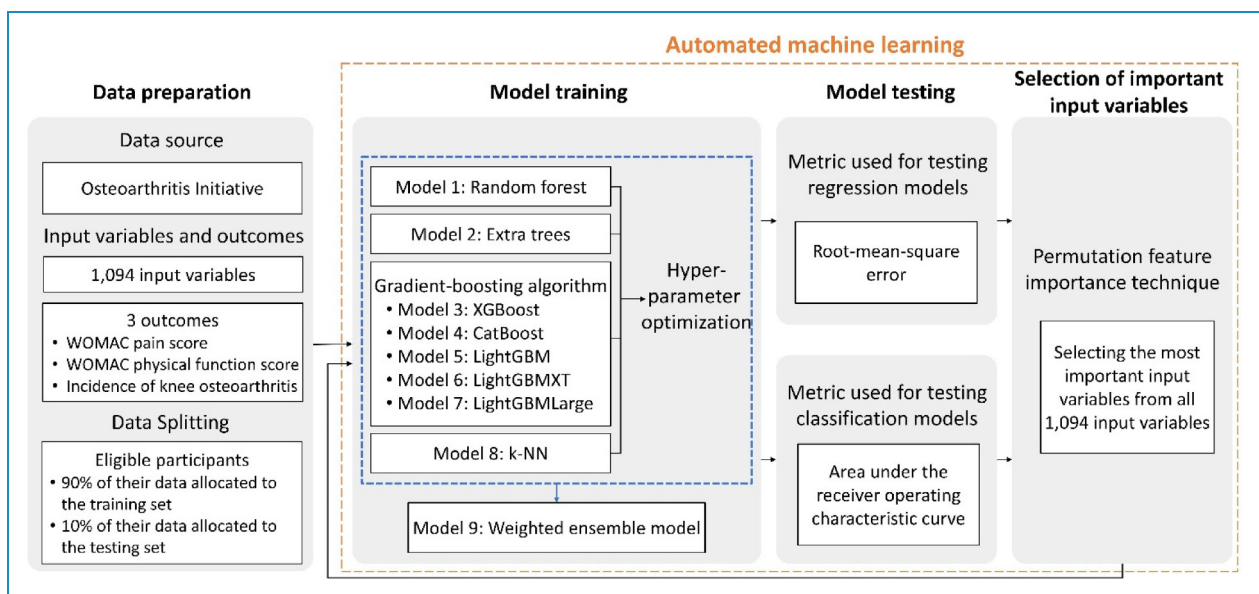

Figure 1. Pipeline of automated machine learning used to predict the progression of knee pain, functional decline, and incidence of knee osteoarthritis in the participants.

in predicting knee pain and functional decline, with a small RMSE indicating high accuracy, that is, a small difference between the predicted and actual WOMAC scores. To determine the incidence of knee OA, we calculated the area under the receiver operating characteristic curve (AUC) to test the accuracy of binary classification, with a large AUC suggesting that a model exhibited good performance in distinguishing between the knee OA and no knee OA conditions. In theory, the weighted ensemble model should perform as good as the best individual models in the training set; however, because of the overfitting issue, it can produce less accurate predictions in the testing set. Thus, we compared the performance of all models to identify any potential differences.

Stage 4—Selection of important input variables. After we had predicted the outcomes using all 1,094 input variables, we employed the technique of permutation feature importance[20,21] to calculate the importance score of each variable and thus identify the most important variables. The higher the score of a variable, the more important it was to the predictive accuracy. To be specific, the value of each variable was permuted individually, and the predictive accuracy was assessed after each permutation; the decrease in the accuracy indicated how much the predictive performance relied on the variable.[21] This study focused on the importance of features in determining the predictive performance of ML models, regardless of the direction of the feature effect. In this study, we identified the five most important variables for the prediction of each outcome in each year.

Results

Study sample

After removing the 84 individuals who had been enrolled in the OAI study but later found to have no outcome data due to knee replacement surgery or confirmed death, a total of 3,200 individuals who were considered at high risk of knee OA at baseline were included as participants in the current study to generate predictive models. Table 1 summarizes the baseline characteristics of the participants.

Prediction results

Figure 2 presents the performances of the ML models for each outcome, and Appendix A shows their detailed performance data. With increasing years, the models' RMSEs increased and their AUCs slightly decreased, indicating that their performances were worsening over time.

Specifically, the results indicated that the weighted ensemble model, category gradient-boosting (CatBoost) model, and extra-trees model exhibited the best performance in predicting pain around the right knee, as these models each had mean RMSEs over nine years of 2.27 (SD = 0.16) on the

Table 1. Baseline characteristics of the 3200 participants in this study.

Characteristics	Mean ± SD or %
Age, years	61.29 ± 9.19
Female	58.97
Body mass index (BMI), kg/m^2	28.10 ± 4.65
Obesity, defined as BMI ≥ 30 kg/m^2	32.81
Presence of knee symptoms	25.72
Previous knee injury	40.22
Previous knee surgery	18.38
Family history of total knee replacement	15.56
Heberden's node in the hand	32.59
Repetitive knee-bending activities	73.06
WOMAC pain (0–20)	
Right knee	1.76 ± 2.58
Left knee	1.65 ± 2.76
WOMAC physical function (0–68)	
Right knee	5.58 ± 8.31
Left knee	5.71 ± 9.25

WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index.

0–20 WOMAC scale. In contrast, we found that the weighted ensemble model showed the best performance for predicting pain around the left knee (mean RMSE = 2.30, SD = 0.13). However, the *k*-nearest neighbors (*k*-NN) model performed the worst in predicting the progression of knee pain over nine years in both knees (right knee: mean RMSE = 2.83, SD = 0.16; left knee: mean RMSE = 2.88, SD = 0.14).

With respect to functional decline over nine years, for the right knee, the weighted ensemble model (mean RMSE = 7.01, SD = 0.65) and CatBoost model (mean RMSE = 7.01, SD = 0.66) showed the best predictive performance, and for the left knee, the weighted ensemble model showed the best predictive performance (mean RMSE = 7.33, SD = 0.59). In contrast, the *k*-NN model showed the worst predictive performance for both knees (right knee: mean RMSE = 9.05, SD = 0.47; left knee: mean RMSE = 9.46, SD = 0.61).

Most of the models exhibited high performance in predicting the incidence of OA in both the right and left knees, with the mean AUC over nine years ranging from
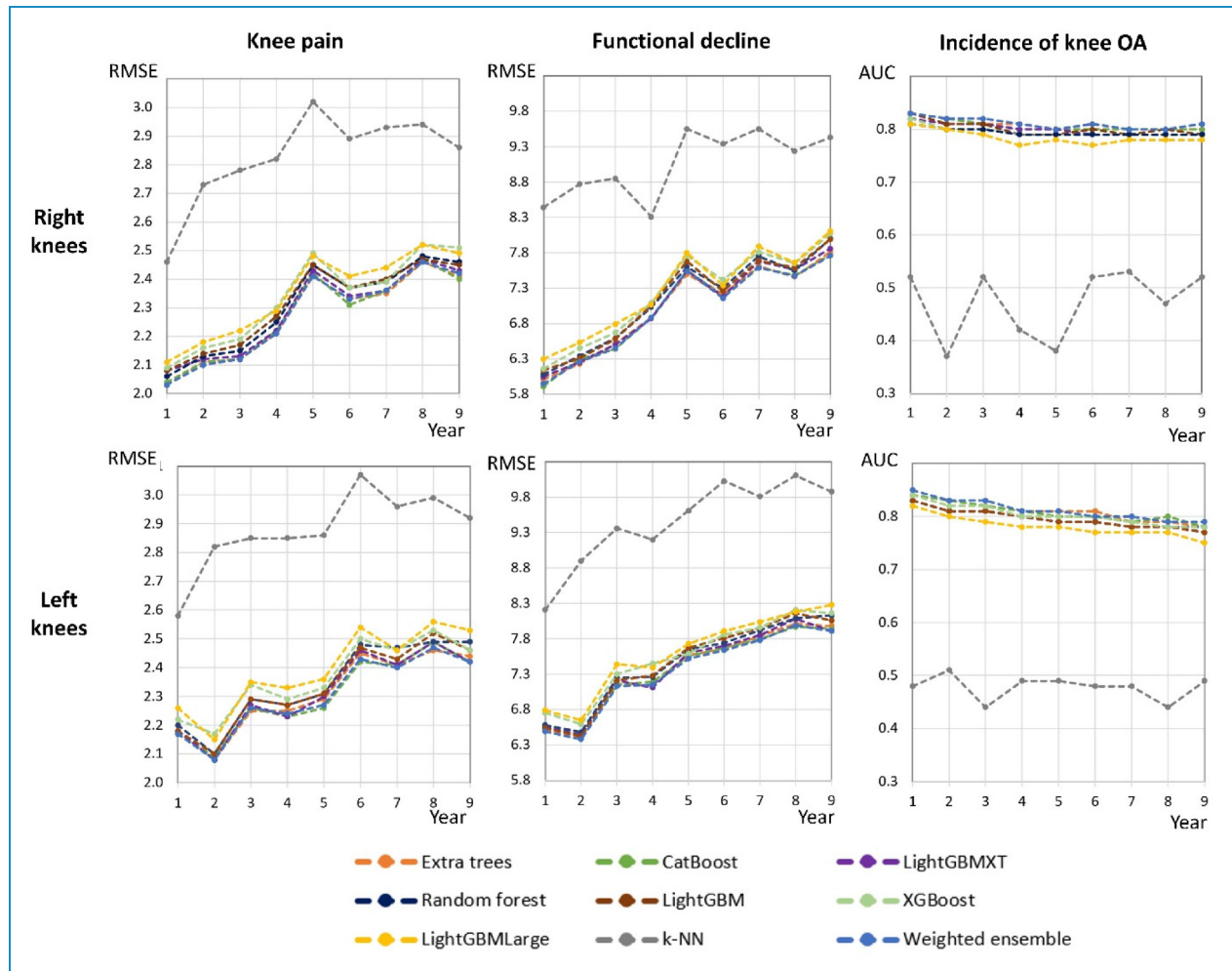
Figure 2. Predictive performances of the models over nine years.

0.78 to 0.81. The best-performing models were the weighted ensemble model, CatBoost model, and extra-trees model (right knee: mean AUC = 0.81, SD = 0.01; left knee: mean AUC = 0.81, SD = 0.02). In contrast, the worst-performing model was the k-NN model, with a mean AUC less than 0.5 for both knees (right knee: mean AUC = 0.47, SD = 0.07; left knee: mean AUC = 0.48, SD = 0.02).

Important input variables for prediction

From all the input variables collected at the OAI baseline visit, we identified the five most important ones for the prediction of each outcome in each year (see Appendix B for the priorities and importance scores of each important variable). The important variables identified for the prediction of all three outcomes over nine years are summarized in Table 2 and were found by analyzing self-reported questionnaire surveys, performance measures, clinical examination reports, and radiographic imaging reports. Some questionnaire-based variables were found to be important

for predicting the progression of knee pain and functional decline, such as the WOMAC score, Knee injury and Osteoarthritis Outcome Score (KOOS), severity of knee pain, and Medical Outcomes Study 12-item Short-Form Health Survey score. With respect to the incidence of knee OA, we found that the radiographic data (including data on the composite OA grade, osteophytes, and joint space narrowing) showed the greatest impacts on prediction results, and other important variables included BMI, knee examination results, dietary nutrition intake, etc.

Discussion

In this study, we applied AutoML to automate the process of using ML models to predict the progression of knee pain, functional decline, and incidence of knee OA in individuals at high risk of such outcomes, and determine the predictive performance of the models. Out of the nine predictive models we used, the weighted ensemble model and the CatBoost model performed the best. This may be

Table 2. Summary of the most important input variables for the prediction of progression of knee pain, functional decline, and incidence of knee OA over nine years.

Important input variables	Progression of knee pain		Functional decline		Incidence of knee OA	
	Right	Left	Right	Left	Right	Left
Age						√
Body mass index			√		√	√
Year of age at first knee injury (left knee)						√
Knee symptoms and knee-related physical function						
WOMAC score (right knee)	√	√	√	√		
WOMAC score (left knee)		√	√	√		
KOOS (right knee)	√	√	√			
KOOS (left knee)		√		√		√
Severity of knee pain on 11-point scale (right knee)	√		√			
Severity of knee pain on 11-point scale (left knee)		√				
Pain: started about how many years ago (left knee)		√				
Other joint symptoms						
Bony enlargement (right hand)						√
Presence of back pain	√					
Presence of left ankle pain				√		
General health						
SF-12 score	√	√	√	√		
Physical activity						
Frequent knee-bending activities: climbing up stairs					√	
Dietary nutrition intake						
Block Brief 2000 Food Frequency Questionnaire		√		√	√	√
Performance measures						
400-meter walk	√					
Knee extension (left knee)	√					
Knee flexion (left knee)						√

Table 2. Continued.

Important input variables	Progression of knee pain		Functional decline		Incidence of knee OA	
	Right	Left	Right	Left	Right	Left
Isometric strength (right leg)						√
Knee examination						
Flexion contracture/hyperextension (right knee)					√	√
Flexion contracture/hyperextension (left knee)						√
Medial tibiofemoral pain (left knee)		√				
Patellofemoral crepitus (right knee)					√	
Radiographic imaging						
Composite OA grade (right knee)	√				√	
Composite OA grade (left knee)						√
Osteophytes and JSN (right knee)					√	
Osteophytes and JSN (left knee)		√	√		√	√

OA = osteoarthritis, WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index, KOOS = Knee injury and Osteoarthritis Outcome Score, SF-12 = Medical Outcomes Study (MOS) 12-item short-form health survey, JSN = joint space narrowing.

attributable to the fact that the weighted ensemble model uses a multi-layer strategy that integrates multiple ML models; as a result, it has demonstrated superior performance in model training and achieved one of the best performances in model testing.[15] It may also be attributable to the fact that compared with traditional gradient-boosting models, the CatBoost model deals better with categorical variables and overcoming the overfitting problem.[22] In contrast, we found that the k-NN model exhibited worse predictive performance than the other models. This may be attributable to the k-NN model having limitations in dealing with a large number of input variables, as it is ineffective at determining the most important variables.[23]

Future studies can adopt our prediction and AutoML approach to automate the process of model training and testing, and then obtain prediction results for clinical use in various contexts. For instance, our approach can be used to predict various outcomes, such as lower limb muscle strength, joint space width, and the probability of total knee replacement. It can also be applied to predictions that use different input variables, including a set of variables predetermined by clinicians and variables collected over multiple years. Furthermore, it can also be used to conduct predictions on different datasets, such as data collected from different patient populations or specific subgroups.

From all input variables collected at baseline, we examined their importance and identified the most important ones for the predictive accuracy of ML models. The identification of the most important variables can help simplify the process of prediction, allowing clinicians to make predictions using a smaller set of important input variables. For the three outcomes investigated in our study, the most important variables varied. It was not surprising to find that the most important variables for predicting the progression of knee pain and functional decline included the subscale and total scores of questionnaire-based measures (e.g. WOMAC and KOOS) collected at baseline, as in our predictions these two outcomes were determined by examining the changes in WOMAC scores.[24] Regarding the prediction of the incidence of knee OA, radiographic data collected at baseline were found to have greater importance than other variables. One reason for this could be that radiographic evidence is considered a mainstay in the diagnosis of knee OA.[25] Another reason could be that the individuals with radiographic evidence of knee OA at baseline were likely to develop knee symptoms,[26,27] which would result in a diagnosis of knee OA with both radiographic evidence and knee symptoms in the follow-up visits.

Moreover, there were differences in important variables for predicting the outcomes in left and right knees. Specifically, when predicting outcomes for the left knees, we found that a majority of important input variables were associated with the left knees; and the same findings were observed for the predictions on right knees. This could be because the data of input variables of one side of the knee can provide a more accurate reflection of the health condition of that particular side, thereby being more influential in predicting outcomes for that corresponding side.

Our findings have several implications for research. First, there was a decreasing trend in the predictive accuracy of our models over time, indicating that short-term prediction was more accurate than long-term prediction. Thus, it is necessary to develop models that can achieve higher accuracy for long-term prediction than the models used in our study. Alternatively, to improve the accuracy, it may be beneficial to use the data from multiple visits as input variables, instead of solely from the baseline visit, as research has shown that using data from longer periods can obtain more robust results compared to using short-term data when predicting the cartilage loss trajectory among OAI participants.[28,29] Second, out of over 1,000 variables, we determined only the five most important input variables for prediction. In future research, it would be useful to allow clinicians to determine the appropriate number of input variables for prediction and select the variables based on clinical need. Third, future studies could use more datasets than we used in this study, as this would enable the validation of AutoML's predictive performance in more healthcare settings, such as predicting the progression of other chronic diseases.

In practical settings, accurate prediction results can enable both clinicians and individuals at high risk to have a clear prospective understanding of the potential progression of knee OA-related symptoms, allowing them to take action to manage knee health in a timely manner before their condition worsens. For example, we performed predictions for both left and right knees to provide more personalized results. This would allow individuals to be aware of the knee that is more vulnerable and take appropriate precautions, such as avoiding injury and minimizing excessive usage of that knee. Moreover, based on personalized prediction results, interventions can be customized to meet the demand for individualized healthcare. For example, individuals at high risk of knee OA who experience a significant decline in physical function could be encouraged to take immediate actions, such as having regular check-ups, implementing body weight control, or taking exercise.[30,31] It is also essential to deliver personalized education to individuals at high risk on the process of OA, its pain mechanisms, and preventive strategies.[32] Further, new interventions can be developed and implemented to improve knee health. Similar to the implementation of other healthcare technologies, their usability, acceptance, and effects on knee health are worth further investigation among the end users.[33–39]

Conclusions

The use of AutoML was found to have the ability to automate the process of using ML models to predict the progression of knee pain, functional decline, and incidence of knee OA in individuals at high risk of these outcomes. Also, the weighted ensemble model and the CatBoost modelshowed the best performance among all nine ML models examined. Among all the input variables, the questionnaire-based outcomes and radiographic data were found to be more important than others in prediction. For future practice, with the use of AutoML, clinicians would be able to generate predictions without the need for extensive knowledge of and experience in ML. This would allow the development of personalized interventions for the prevention of health issues. Moreover, the use of this predictive approach for assessing individuals at high risk of knee OA would help them to be made sufficiently aware of their knee problems that they would take immediate action to improve their knee health.

ORCID iD: Calvin Kalun Or https://orcid.org/0000-0002-9819-8865

References

1. Felson DT, Naimark A, Anderson J, et al. The prevalence of knee osteoarthritis in the elderly. the Framingham osteoarthritis study. *Arthritis Rheum* 1987; 30: 914–918.
2. Jordan JM, Linder GF, Renner JB, et al. The impact of arthritis in rural populations. *Arthritis Rheum* 1995; 8: 242–250.

3. Cui A, Li H, Wang D, et al. Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies. *EClinicalMedicine* 2020; 29: 100587.

4. Palazzo C, Nguyen C, Lefevre-Colau M-M, et al. Risk factors and burden of osteoarthritis. *Ann Phys Rehabil Med* 2016; 59: 134–138.

5. Peat G, McCarney R and Croft P, Knee pain and osteoarthritis in older adults: a review of community burden and current use of primary health care. *Ann Rheum Dis*, 2001; 60: 91–97.

6. Hunter DJ and Bierma-Zeinstra S. Osteoarthritis. *Lancet* 2019; 393: 1745–1759.

7. Brahim A, Jennane R, Riad R, et al. A decision support tool for early detection of knee osteoarthritis using X-ray imaging and machine learning: data from the Osteoarthritis Initiative. *Comput Med Imaging Graph* 2019; 73: 11–18.

8. Buckland-Wright C. Which radiographic techniques should we use for research and clinical practice? *Best Pract Res Clin Rheumatol* 2006; 20: 39–55.

9. Guermazi A, Hayashi D, Roemer FW, et al. Osteoarthritis: a review of strengths and weaknesses of different imaging options. *Rheum Dis Clin* 2013; 39: 567–591.

10. Zhang W, McWilliams DF, Ingham SL, et al. Nottingham knee osteoarthritis risk prediction models. *Ann Rheum Dis* 2011; 70: 1599–1604.

11. Long MJ, Papi E, Duffell LD, et al. Predicting knee osteoarthritis risk in injured populations. *Clin Biomech* 2017; 47: 87–95.

12. Kotthoff L, Thornton C, Hoos HH, et al. Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA, in *Automated machine learning: methods, systems, challenges.* 2019, Springer: Cham.

13. Jamshidi A, Leclercq M, Labbe A, et al., Identification of the most important features of knee osteoarthritis structural progressors using machine learning methods. *Ther Adv Musculoskelet Dis* 2020; 12: 1–12.

14. Waring J, Lindvall C and Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med* 2020; 104: 101822.

15. Erickson N, Mueller J, Shirkov A, et al. AutoGluon-tabular: Robust and accurate autoML for structured data. *arXiv*, 2020.

16. He X, Zhao K and Chu X. AutoML: a survey of the state-of-the-art. *Knowl Based Syst* 2021; 212: 106622.

17. Eckstein F, Wirth W and Nevitt MC, Recent advances in osteoarthritis imaging—the osteoarthritis initiative. *Nat Rev Rheumatol* 2012; 8: 622–630.

18. Bellamy N, Buchanan WW, Goldsmith CH, et al. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988; 15: 1833–1840.

19. Cawley GC and Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010; 11: 2079–2107.

20. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.

21. Aldrich C and Auret L, Fault detection and diagnosis with random forest feature extraction and variable importance methods. *IFAC Proc Vol* 2010; 43: 79–86.

22. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: Unbiased boosting with categorical features. *arXiv*, 2017.

23. Domingos P. A few useful things to know about machine learning. *Commun ACM* 2012; 55: 78–87.

24. Halilaj E, Le Y, Hicks JL, et al. Modeling and predicting osteoarthritis progression: data from the Osteoarthritis Initiative. *Osteoarthritis Cartilage* 2018; 26: 1643–1650.

25. Kohn MD, Sassoon AA and Fernando ND, Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. *Clin Orthop Relat Res* 2016; 474: 1886–1893.

26. Neogi T, Frey-Law L, Scholz J, et al. Sensitivity and sensitisation in relation to pain severity in knee osteoarthritis: trait or state? *Ann Rheum Dis* 2015; 74: 682–688.

27. Neogi T, The epidemiology and impact of pain in osteoarthritis. *Osteoarthritis Cartilage* 2013; 21: 1145–1153.

28. Eckstein F, Boudreau RM, Wang Z, et al. Trajectory of cartilage loss within 4 years of knee replacement–a nested case-control study from the osteoarthritis initiative. *Osteoarthritis Cartilage* 2014; 22 1542–1549.

29. Eckstein F, Mc Culloch CE, Lynch JA, et al. How do short-term rates of femorotibial cartilage change compare to long-term changes? Four year follow-up data from the osteoarthritis initiative. *Osteoarthritis Cartilage* 2012; 20: 1250–1257.

30. Roos EM and Arden NK, Strategies for the prevention of knee osteoarthritis. *Nat Rev Rheumatol* 2016; 12: 92–101.

31. Chen T, Or CK and Chen J. Effects of technology-supported exercise programs on the knee pain, physical function, and quality of life of individuals with knee osteoarthritis and/or chronic knee pain: a systematic review and meta-analysis of randomized controlled trials. *J Am Med Inform Assoc* 2021; 28: 414–423.

32. Bennell KL, Hunter DJ and Hinman RS. Management of osteoarthritis of the knee. *Br Med J* 2012; 345: e4934.

33. Chen T, Chen J, Or CK, et al. Path analysis of the roles of age, self-efficacy, and TAM constructs in the acceptance of performing upper limb exercises through immersive virtual reality games. *Int J Ind Ergon* 2022; 91: 103360.

34. Chen T and Or CK. Perceptions of a machine learning-based lower-limb exercise training system among older adults with knee pain. *Digital Health* 2023; 9: 1–11.

35. Chen T and Or CK. Development and pilot test of a machine learning-based knee exercise system with video demonstration, real-time feedback, and exercise performance score. In: *Proceedings of the human factors and ergonomics society annual meeting*, 2021. Los Angeles, CA: Sage Publications Sage CA, pp.1519–1523.

36. Chen J, Or CK and Chen T. Effectiveness of using virtual reality–supported exercise therapy for upper extremity motor rehabilitation in patients with stroke: systematic review and meta-analysis of randomized controlled trials. *J Med Internet Res* 2022; 24: e24111.

37. Chen T, Or CK and Chen J. A systematic review and meta-analysis of randomized controlled trials to evaluate technology-supported exercise programs for knee health. In: *Proceedings of the human factors and ergonomics society annual meeting*, 2020. Sage CA: Los Angeles, CA, pp.639–640.

38. Guo J, Chen T, Xie Z, et al. Effects of interventions to reduce the negative consequences of interruptions on task performance: a systematic review, meta-analysis, and narrative synthesis of laboratory studies. *Appl Ergon* 2021; 97: 103506.

39. Chen T, Guo J, Xie Z, et al. Using a checklist-based electronic calculation aid and contextual cues to assist medication dose calculation and calculation resumption after interruptions. *Int J Ind Ergon* 2023; 95: 103453.