

Methods of Information in Medicine

Deep Learning for Predicting Progression of Patellofemoral Osteoarthritis Based on Lateral Knee Radiographs, Demographic Data and Symptomatic Assessments

Neslihan Bayramoglu, Martin Englund, Ida K Haugen, Muneaki Ishijima, Simo Saarakkala.

Affiliations below.

DOI: 10.1055/a-2305-2115

Please cite this article as: Bayramoglu N, Englund M, Haugen I K et al. Deep Learning for Predicting Progression of Patellofemoral Osteoarthritis Based on Lateral Knee Radiographs, Demographic Data and Symptomatic Assessments. *Methods of Information in Medicine* 2024. doi: 10.1055/a-2305-2115

Conflict of Interest: The authors declare that they have no conflict of interest.

Abstract:

Objective: In this study, we propose a novel framework that utilizes deep learning and attention mechanisms to predict the radiographic progression of patellofemoral osteoarthritis (PFOA) over a period of seven years.

Design: This study included subjects (1832 subjects, 3276 knees) from the baseline of the Multicenter Osteoarthritis Study (MOST). Patellofemoral joint regions-of-interest were identified using an automated landmark detection tool (BoneFinder) on lateral knee X-rays. An end-to-end deep learning method was developed for predicting PFOA progression based on imaging data in a 5-fold cross-validation setting. To evaluate the performance of the models, a set of baselines based on known risk factors were developed and analyzed using gradient boosting machine (GBM). Risk factors included age, sex, BMI and WOMAC score, and the radiographic osteoarthritis stage of the tibiofemoral joint (KL score). Finally, to increase predictive power, we trained an ensemble model using both imaging and clinical data.

Results: Among the individual models, the performance of our deep convolutional neural network attention model achieved the best performance with an AUC of 0.856 and AP of 0.431; slightly outperforming the deep learning approach without attention (AUC=0.832, AP= 0.4) and the best performing reference GBM model (AUC=0.767, AP= 0.334). The inclusion of imaging data and clinical variables in an ensemble model allowed statistically more powerful prediction of PFOA progression (AUC = 0.865, AP=0.447), although the clinical significance of this minor performance gain remains unknown. The spatial attention module improved the predictive performance of the backbone model, and the visual interpretation of attention maps focused on the joint space and the regions where osteophytes typically occur.

Conclusion: This study demonstrated the potential of machine learning models to predict the progression of PFOA using imaging and clinical variables. These models could be used to identify patients who are at high risk of progression and prioritize them for new treatments. However, even though the accuracy of the models were excellent in this study using the MOST dataset, they should be still validated using external patient cohorts in the future.

Corresponding Author:

Dr. Neslihan Bayramoglu, University of Oulu, Research Unit of Health Sciences and Technology, Oulu, Finland, neslihan.bayramoglu@oulu.fi

Affiliations:

Neslihan Bayramoglu, University of Oulu, Research Unit of Health Sciences and Technology, Oulu, Finland



This article is protected by copyright. All rights reserved.

Accepted Manuscript

Deep Learning for Predicting Progression of Patellofemoral Osteoarthritis Based on Lateral Knee Radiographs, Demographic Data and Symptomatic Assessments

Neslihan Bayramoglu^{a*}, Martin Englund^b, Ida K. Haugen^c, Muneaki Ishijima^d, Simo Saarakkala^{a,e}

^a*Research Unit of Health Sciences and Technology, University of Oulu, Finland*

^b*Orthopaedics, Department of Clinical Sciences Lund Faculty of Medicine, Lund University*

^c*Center for Treatment of Rheumatic and Musculoskeletal Diseases (REMEDY), Diakonhjemmet Hospital, Oslo, Norway*

^d*Department of Orthopaedics, Faculty of Medicine, Juntendo University*

^e*Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland*

*

Corresponding author

Neslihan Bayramoglu, Ph.D., Docent, Research Unit of Health Sciences and Technology, University of Oulu, Finland

neslihan.bayramoglu@oulu.fi

POB 5000, FI-90014, Oulu, Finland

Abstract

Objective: In this study, we propose a novel framework that utilizes deep learning and attention mechanisms to predict the radiographic progression of patellofemoral osteoarthritis (PFOA) over a period of seven years.

Design: This study included subjects (1832 subjects, 3276 knees) from the baseline of the Multicenter Osteoarthritis Study (MOST). Patellofemoral joint regions-of interest were identified using an automated landmark detection tool (BoneFinder) on lateral knee X-rays. An end-to-end deep learning method was developed for predicting PFOA progression based on imaging data in a 5-fold cross-validation setting. To evaluate the performance of the models, a set of baselines based on known risk factors were developed and analyzed using gradient boosting machine (GBM). Risk factors included age, sex, BMI and WOMAC score, and the radiographic osteoarthritis stage of the tibiofemoral joint (KL score). Finally, to increase predictive power, we trained an ensemble model using both imaging and clinical data.

Results: Among the individual models, the performance of our deep convolutional neural network attention model achieved the best performance with an AUC of 0.856 and AP of 0.431; slightly outperforming the deep learning approach without attention (AUC=0.832, AP= 0.4) and the best performing reference GBM model (AUC=0.767, AP= 0.334). The inclusion of imaging data and clinical variables in an ensemble model allowed statistically more powerful prediction of PFOA progression (AUC = 0.865, AP=0.447), although the clinical significance of this minor performance gain remains unknown. The spatial attention module improved the predictive performance of the backbone model, and the visual interpretation of attention maps focused on the joint space and the regions where osteophytes typically occur.

Conclusion: This study demonstrated the potential of machine learning models to predict the progression of PFOA using imaging and clinical variables. These models

could be used to identify patients who are at high risk of progression and prioritize them for new treatments. However, even though the accuracy of the models were excellent in this study using the MOST dataset, they should be still validated using external patient cohorts in the future.

Keywords: Patellofemoral Osteoarthritis, Deep Learning, Prediction of Osteoarthritis Progression, Knee

1. Introduction

Knee osteoarthritis (OA) is the most prevalent chronic joint disorder that involves degeneration and loss of articular cartilage along with bony changes. High age and body mass index (BMI) are strong risk factors for knee OA¹. Structural knee OA often leads to significant pain, stiffness, disability, and reduced quality of life for affected individuals². Current understanding of OA disease process is inadequate and, consequently, there is a lack of disease modifying medical treatments. As a result, knee OA continues to impose a significant burden on individuals and society³.

Although the patellofemoral (PF) joint is an important source of symptoms in knee OA, the majority of the research on knee OA has focused on tibiofemoral (TF) joint of the knee^{4,5}. Patellofemoral OA (PFOA) can be caused by a number of factors, including previous injury to the knee, inflammation, biomechanical abnormalities, overuse of joint, obesity, and genetic predisposition^{3,6}. Symptoms often include anterior knee pain, especially when kneeling and squatting, as well as swelling and a grinding or popping sensation when moving the knee (crepitus)⁷. As the importance of the PF joint in OA is increasingly acknowledged, the number of studies into it has been increasing^{3,8,9}. Still, more research is needed³.

Non-invasive imaging techniques play a crucial role in diagnosing and monitoring PFOA. Without imaging, a confident diagnosis will seldom be possible for PFOA¹⁰. X-ray imaging is one of the primary diagnostic tools because of its low cost and wide availability. Although radiography does not allow to visualize soft tissues, changes in the joint space and bone structure can be well depicted from X-rays. Several imaging biomarkers such as the narrowing of the joint space, bony spurs, malalignment of the patella, bone sclerosis, and cysts are associated with PFOA^{9,11}.

In recent years, machine learning (ML) techniques have emerged as promising tools to aid in the diagnosis of PFOA from X-ray images^{12,13}. Both early diagnosis and prediction of disease progression might be critical in the management and intervention of PFOA. However, accurate and timely identification of PFOA progression based on X-ray images can be challenging due to the complexity of the disease and the variability of knee imaging. To date, there are no published studies using ML-based models for prediction of PFOA development or progression in the future from imaging data.

In this study, we introduced a deep learning based framework to predict radiographic progression of PFOA over a 7-year period from lateral radiographs, demographic data and symptom assessments (clinical data). We leveraged attention mechanism in our deep learning framework and proposed an end-to-end solution via a trainable attention module. The results of this study have the potential to improve the early diagnosis and treatment of PFOA, ultimately leading to improved patient outcomes and quality of life.

2. Materials and Methods

Figure 1 shows the overall pipeline of our study. We first located patellar landmarks using BoneFinder software¹⁴ (Figure 4). Those anatomical landmarks were then used to align patellar bone constantly across the knees eliminating rotation variance.

The image preprocessing step involved normalizing intensity using global contrast normalization and truncating the histogram between the 5th and 99th percentiles. Subsequently, we used patellar landmarks to locate the patellofemoral joint regions of interest (PFJROI) in lateral knee radiographs. To ensure a similar view with left knee images, the right knee ROI images were horizontally flipped. We then utilized a deep convolutional neural network (CNN) to predict PFOA progression within 7 years. Additionally, we trained a machine learning model (GBM¹⁵) on clinical features as a reference method for comparison with the proposed approach. Finally, to increase predictive power, we trained an ensemble model using both imaging and clinical data.

2.1. Data

We used the data from the Multicenter Osteoarthritis Study public use datasets (MOST, <http://most.ucsf.edu>). MOST is a longitudinal observational study that aims to identify factors affecting the occurrence and progression of OA. The study enrolled 3,026 participants aged 50–79 years who either had radiographic knee OA or were at high risk for developing the disease. The participants has been followed 84 months where clinical assessments were conducted and radiological data were collected. In the study, semiflexed lateral view radiographs were acquired according to a standardized protocol. Knee radiographs were evaluated from the baseline to 15, 30, 60 and 84-month follow-up visits. In this study, we employed lateral radiographs acquired at the baseline visit from both left and right legs that includes 3276 knees (1832 subjects) which did not have PFOA at the time of first examination. The number of progressed knees that developed PFOA is 403 (12%) and the number of knees which did not develop PFOA is 2873 (88%). Selected knees must have had PFOA assessments from lateral radiographs and KL grades from posteroanterior (PA) radiographs, all performed at the baseline. Among those ones, we selected knees only whose patellofemoral OA

status within the following 7 years can be assessed (progressor vs non-progressor). For example, participants who dropped out from the study before the last follow-up timepoint and had not developed PFOA at the previous time points were excluded. See Figure 2 and Table 1 for subject flow diagram and demographics.

In the MOST public use datasets, radiographic PFOA is defined from lateral view radiographs as follows: Osteophyte score ≥ 2 or the joint space narrowing (JSN) score is ≥ 1 plus any osteophyte, sclerosis or cysts ≥ 1 in the PF joint (grades 0–3; 0=normal, 1=mild, 2=moderate, 3=severe). Unlike tibiofemoral joint OA assessment (KL grading ranging from 0 to 4), in the PF joint, OA was described either present or absent lacking a severity grading. In this study, the term “progression” refers to both progression of existing OA and development of OA in previously non-affected PF joints (incidence). For example, knees which showed minor signs of PFOA (e.g. osteophyte score=1) at the baseline, which are still considered as non PFOA cases, might experience worsening of an existing abnormality in the following years and diagnosed with PFOA (progression). Similarly, knees that did not show any signs of PFOA at the baseline might develop the disease for the first time during the the following 7 years (incidence). In MOST, individual radiographic features were graded by two independent expert readers and when there was a disagreement in film readings, a panel of three adjudicators resolved the discrepancies¹⁶.

2.2. Selection of Region of Interest (ROI)

We placed a PFJROI automatically using landmarks (Figure 4). The height of the patellar bone (h) was used to locate a square shaped image ROI. Once the patellar bone margins were determined using landmarks, a 20 pixels $\left(\frac{\Delta}{2}\right)$ region is padded around the bone. On the femur side, the ROI is extended to capture the part of the femur facing the patellar bone such that the width of the ROI equals to the height of it

($height = width = h + \Delta$). Finally, the size of the ROI becomes proportional to the size of the patellar bone.

2.3. Predicting progression of patellofemoral osteoarthritis using Deep CNN

We adopted the deep CNN architecture proposed by Yan et al.¹⁷ to predict PFOA development based on the baseline imaging data. It uses VGG-16¹⁸ backbone with two additional attention layers and one penultimate global feature vector (obtained via global average pooling)(Figure 1). PFJROI data were pre-processed by resizing it to 256×256 pixels and then applying a random crop of size 224×224 pixels. The backbone network VGG-16 was initialized with its pre-trained version on ImageNet. The attention modules were initialized using He's initialization¹⁹. We employed Focal loss²⁰, a variant of the cross-entropy, which has shown to be an effective when facing the class imbalance problem by selectively downweighting well-classified examples. We used a batch size of 32 and trained the network end-to-end for 45 epochs using stochastic gradient descent with momentum. The initial learning rate was 0.001 and it was decayed by 10 every 10 epochs.

To examine the impact of the attention mechanism on the model's performance, a separate training was conducted with the original VGG-16 network without the attention modules. The network parameters were initialized with ImageNet pre-training, and the last layer was modified for binary classification. To ensure a fair comparison, we maintained consistency in the other network parameters and hyper-parameters between the attention model and the model without attention.

2.4. Attention Module

Previous deep learning works that employ post hoc analysis for visual explanations such as Grad-CAM²¹ require extra computation based on a fully trained classification

network and relies on gradient information passed to the last convolutional layer combined with the forward activation maps. However, those feature maps, that are often used to produce explanations, are not necessarily related to the target class and they do not affect the network parameters at all. In this study, we employed a trainable spatial attention mechanism to produce insights into the model decisions. Attention mechanisms are widely used in the field of natural language processing (NLP) as a way to improve the performances of models by emphasizing the important parts of the information²². In case of image classification, the idea of trainable attention is to focus on the most informative parts of an image while ignoring less relevant or noisy parts. During training, the network learns to weight different regions of the input image based on the classification performance. See Supplementary Material for more details of the attention module used in our architecture.

2.5. Reference Models

We employed GBM to predict the development of PFOA from demographic data and self-reported symptom assessments. GBM is a popular and powerful machine learning algorithm used for regression and classification based on ensembles of decision trees¹⁵. It works iteratively by adding decision trees to the model where each new tree attempts to correct the errors made by the previous trees. In this study, we used an efficient implementation of GBM called LightGBM²³.

We built three GBM classifiers based on the clinical data and risk factors. These include age, sex, body-mass index (BMI), the total Western Ontario and McMaster Universities Arthritis Index (WOMAC) score, and the KL grade of the tibiofemoral joint (Model1, Model2, and Model3 in Table 2). The WOMAC score is a widely used questionnaire-based assessment tool designed to evaluate the severity of pain, stiffness, and physical disability in patients with OA of the knee and hip.

For all of our models, we utilized subject-wise stratified 5-fold cross validation. This involves dividing the dataset into 5 folds, each containing data from different subjects, and stratifying the data within each fold so that the proportion of progressors vs non-progressors is similar to the overall dataset. This helps to eliminate subject-dependent bias between the training and validation sets.

K-fold cross-validation involves iteratively selecting one fold as the testing set and the remaining folds as the training set. The model is trained on the training set and evaluated on the testing set. This process is repeated for each fold, with each fold serving as the testing set exactly once.

To ensure fair comparisons, we used the same folds for all of the models. All of the models were trained separately and the reported performances were derived from these separate models.

2.6. Statistical Methods

The performance of the models were compared using Receiver Operating Characteristics (ROC) curves, Precision-Recall (PR) curves, and Brier score²⁴. ROC curves plot the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. The area under the ROC curve (AUC-ROC) is often used as a summary metric for model performance, with a value of 1 indicating perfect classification and 0.5 indicating random classification. On the other hand, PR curves plot the precision (positive predictive value) against the recall (true positive rate) at various classification thresholds. The area under the PR curve (average precision, AP) is another commonly used summary metric for model performance, with a value of 1 indicating perfect classification and 0 indicating random classification. ROC curves are often used when the number of negative instances is much larger than the number of positive instances, while PR curves are more suitable when the number of positive

instances is relatively small. In general, a good classifier should have high values for both AUC-ROC and AUC-PR. Brier score equals to the mean squared error of the prediction. In order to compare the differences between model AUCs, we applied DeLong's test²⁵.

3. Results

Table 2 and Figure 5 show the performance of different models in predicting PFOA progression. Our proposed VGG-16-Attn model achieved the highest AUC of 0.856 [0.838, 0.872] and AP of 0.431 [0.387, 0.475] among all the considered models (Model1 to Model5). We compared the performance of VGG-16-Attn with the original VGG-16 model to assess the contribution of attention modules. Our results show that the addition of attention modules has a positive impact on the performance of the model, with a statistically significant difference between the AUC values of the two models (DeLong's p-value = 0.00018).

To assess the value of imaging biomarkers in predicting PFOA progression, we conducted a thorough evaluation of various risk factors, including age, sex, body-mass index (BMI), WOMAC, and TFOA KL scores (Figure 5) as reference models. Using gradient boosting machine (GBM) models, we trained the models to predict the probability of developing PFOA based on different combinations of these risk factors. Our results showed that the best-performing reference model (Model3) incorporated age, sex, BMI, WOMAC, and TFOA KL scores, achieving an AUC of 0.767 [0.74, 0.789] and an AP of 0.334 [0.293, 0.377] (Figure 5). We also measured the impact of each feature on the model's output by looking at the contribution of that feature to the predicted outcome compared to what the predicted outcome would be if the feature was not included in the model (SHapley Additive exPlanations²⁶ (Supplementary Figure S1

and Supplementary Figure S2). High BMI, WOMAC and KL scores increase the predicted PFOA progression risk and low BMI, WOMAC, KL scores reduce the risk.

Subsequently, we compared the performance of our deep convolutional neural network (CNN) attention model (VGG-16-Attn, Model5) to the bestperforming reference method (Model3). Our results showed a statistically significant difference between the AUC values of the two models (DeLong's p -value $< 1e - 10$).

To further improve predictive accuracy, we used a second-layer GBM model that fused the predictions of the VGG-16-Attn CNN model (Model5) and the strongest reference model (Model3) with imaging features and clinical assessments (Figure 1c). This stacked model (Model6) achieved the best AUC of 0.865 [0.838, 0.872], an AP of 0.447 [0.404, 0.491], and a Brier score of 0.084, outperforming both individual models. While the increase in AUC between the stacked model (Model6) and the VGG-16-Attn CNN model (Model5) was statistically significant (DeLong's p -value = 0.0085), it was not highly significant.

Examples of spatial attention maps are presented in Figure 7. The shallower attention map which is applied after conv3 layer, focus on more general and diffused areas. Therefore, we present here only the deeper attention map (after the conv4 layer in Figure 1). In various cases, the model paid attention to the PF joint space width and the inferior and posterior regions of patellar bone. Additional examples of such attention maps are presented in the Supplementary.

4. Discussion

This study presents a novel deep learning-based approach for predicting progression of PFOA, utilizing both clinical variables and imaging data. The results demonstrate the potential of machine learning techniques, especially deep learning, in

predicting PFOA progression, which could provide valuable information for clinicians in patient care.

In general, ML-based models can handle heterogeneous data and they can identify patterns that may not be apparent to human experts. We highlighted this by the inclusion of both clinical variables and imaging data into the stacked model. This combination model achieved the highest accuracy in predicting PFOA progression, indicating its ability to differentiate between patients who are likely to experience PFOA and those who are not. However, it should be still noted that the performance gain with the stacked model (AUC=0.865, AP=0.447), compared to the imaging-based model (AUC=0.856, AP=0.431), was only minor and, although statistically significant, probably the clinical gain might be insignificant. Consequently, this suggests that clinical variables have only minor contribution to the prediction performance on top of the X-ray image alone. Similarly as in the case of knee OA progression prediction²⁷, it looks like that a knee lateral X-ray image already includes indirectly a lot of clinical information, such as age and BMI.

Our study confirmed that high BMI, high WOMAC score, female sex, and OA in the TF joint (KL score) are all risk factors for PFOA development (Supplementary Figures S1 and S2 and (Table 1). Out of the three main demographical variables age, sex and BMI in isolation (Model1), the strongest predictive capability was high BMI.

It has been earlier reported that the use of attention mechanism increases the performance of NLP models^{22,28}. Here, we also observed the increased performance in this kind of image classification task (AUC = 0.856 vs. 0.832, AP = 0.431 vs. 0.400). Besides the increase in overall model performance, generated attention maps highlighted the joint space and the regions where osteophytes typically occur. These regions are known to be affected in PFOA, and they reflect manual imaging biomarkers

of OA including joint space narrowing and morphological and structural changes in bone.

The present study is unique as it investigated the potential of machine learning approaches based on imaging data to accurately predict PFOA progression for the first time. However, there are also some limitations of this study. First and foremost, the model was trained on data from a single population, and further research is necessary to validate the model's generalizability to other populations and settings. Additionally, the study did not consider other potential predictors of PFOA progression, such as biomechanical or genetic factors. Incorporating longitudinal data and other types of imaging data, such as MRI, could further improve the model.

In conclusion, our study demonstrates the potential of machine learning models to predict PFOA progression using imaging and clinical variables. These models could assist in identifying patients at high risk of PFOA progression, enabling clinicians to intervene with personalized treatment plans and potentially prevent or delay disease progression.

Summary

- We present the first study for predicting PFOA progression based on a multi-modal machine learning method using lateral X-ray images and clinical data.
- We leveraged trainable attention mechanism to highlight regions in lateral X-rays which highly contributed the decision of the model.

- We compared the performances of deep convolutional neural network based models and gradient boosting machine based models using clinical variables including age, sex, body mass index (BMI), the total Western Ontario and McMaster Universities Arthritis Index (WOMAC) score, and Kellgren and Lawrence (KL) score of the tibiofemoral joint.
- Finally, we proposed a stacked model where both deep CNN predictions and predictions from clinical model are combined with a second level machine learning model - Gradient Boosting Machine (GBM).
- Our results demonstrated that imaging biomarkers contain useful information for predicting PFOA progression within 7 years. Moreover, addition of clinical data slightly improves the prediction power of the imaging-based model, although the clinical significance of this performance gain is unknown.
- Predicting PFOA progression/ development have the potential to improve the early diagnosis, management and treatment of PFOA.

Acknowledgements

Multicenter Osteoarthritis Study (MOST) Funding Acknowledgment. MOST is comprised of four cooperative grants (Felson – AG18820; Torner – AG18832, Lewis – AG18947, and Nevitt – AG19069) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by MOST study investigators. This manuscript was prepared using MOST data and does not necessarily reflect the opinions or views of MOST investigators.

We would like to acknowledge the NORDFORSK grant from the project “Molecular and structural biomarkers for personalised care in osteoarthritis” (Project No.: 116406).

We gratefully acknowledge Claudia Lindner for providing the BoneFinder® tool and lateral knee active shape model, Aleksei Tiulpin for providing an interface to BoneFinder to fully leverage multiple processors, and the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used in this research.

Funding

Funding sources are not associated with the scientific contents of the study.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

N.B. originated the idea of the study. N.B. performed the experiments and took major part in writing of the manuscript. S.S. supervised the project. All authors participated in producing the final manuscript draft and approved the final submitted version.

References

1. Lankhorst NE, Damen J, Oei EH, et al. Incidence, prevalence, natural course and prognosis of patellofemoral osteoarthritis: the Cohort Hip and Cohort Knee study. *Osteoarthritis and Cartilage*. 2017;25(5):647-653

2. Duncan R, Peat G, Thomas E, Wood L, Hay E, Croft P. How do pain and function vary with compartmental distribution and severity of radiographic knee osteoarthritis? *Rheumatology*. 2008;47:1704–1707
3. Crossley KM, Hinman RS. The patellofemoral joint: the forgotten joint in knee osteoarthritis *Osteoarthritis and Cartilage*. 2011;19:765–767
4. Duncan R, Peat G, Thomas E, Hay EM, Croft P. Incidence, progression and sequence of development of radiographic knee osteoarthritis in a symptomatic population *Annals of the rheumatic diseases*. 2011;70:1944–1948
5. Duncan R, Peat G, Thomas E, Wood L, Hay E, Croft P. Does isolated patellofemoral osteoarthritis matter? *Osteoarthritis and cartilage*. 2009;17:1151–1155
6. Kim YM, Joo YB. Patellofemoral osteoarthritis *Knee Surgery and Related Research*. 2012;24:193–200
7. van Middelkoop M, Bennell KL, Callaghan MJ, et al. International patellofemoral osteoarthritis consortium: consensus statement on the diagnosis, burden, outcome measures, prognosis, risk factors and treatment. In: *Seminars in Arthritis and Rheumatism*. Vol 47. Elsevier; 2018:666-675
8. Kobayashi S, Pappas E, Fransen M, Refshauge K, Simic M. The prevalence of patellofemoral osteoarthritis: a systematic review and metaanalysis *Osteoarthritis and cartilage*. 2016;24:1697–1707

9. Macri EM. Patellofemoral osteoarthritis: characterizing knee alignment and morphology [PhD thesis]. British Columbia, University of British Columbia 2017
10. Peat G, Duncan RC, Wood LRJ, Thomas E, Muller S. Clinical features of symptomatic patellofemoral joint osteoarthritis. *Arthritis research & therapy*. 2012;14(2):R63
11. de Lange-Brokaar BJE, Bijsterbosch J, Kornaat PR, et al. Radiographic progression of knee osteoarthritis is associated with MRI abnormalities in both the patellofemoral and tibiofemoral joint. *Osteoarthritis and cartilage*. 2016;24(3):473-479
12. 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) *Osteoarthritis and Cartilage*. 2021;29:1432–1447
13. Bayramoglu N, Nieminen NT, Saarakkala S. Machine learning based texture analysis of patella from X-rays for detecting patellofemoral osteoarthritis *International journal of medical informatics*. 2022;157:104627
14. Lindner C, Thiagarajah S, Wilkinson JM, et al. Fully automatic segmentation of the proximal femur using random forest regression voting *IEEE transactions on medical imaging*. 2013;32:1462– 1472
15. Friedman JH. Greedy function approximation: a gradient boosting machine *Annals of statistics*. 2001:1189–1232

16. Roemer FW, Guermazi A, Hunter DJ, et al. The association of meniscal damage with joint effusion in persons without radiographic osteoarthritis: the Framingham and MOST osteoarthritis studies *Osteoarthritis and cartilage*. 2009;17:748–753
17. Yan Y, Kawahara J, Hamarneh G. Melanoma recognition via visual attention in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings* 26:793–804 Springer 2019
18. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition arXiv preprint arXiv:1409.1556. 2014
19. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification in *Proceedings of the IEEE international conference on computer vision*:1026–1034 2015
20. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection in *Proceedings of the IEEE international conference on computer vision*:2980–2988 2017
21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization in *Proceedings of the IEEE international conference on computer vision*:618–626 2017
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need *Advances in neural information processing systems*. 2017;30

23. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree in *Advances in neural information processing systems*:3146–3154 2017
24. Brier GW. Verification of forecasts expressed in terms of probability *Monthly weather review*. 1950;78:1–3
25. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach *Biometrics*. 1988:837–845
26. Lundberg SM, Lee Su-In. A Unified Approach to Interpreting Model Predictions in *Advances in Neural Information Processing Systems* 30:4765–4774 2017
27. Tiulpin A, Klein S, Bierma-Zeinstra S, et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Scientific reports*. 2019 Dec 27;9(1):20038
28. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning *Neurocomputing*. 2021;452:48–62

Figure1: a) Illustration of the workflow of our approach. The localization and alignment of patellofemoral (PF) joint in lateral knee X-rays were performed based on the anatomical landmarks of patellar bone (BoneFinder). Intensity normalization was then applied. Finally, each lateral knee was rotated in order to have an aligned patella. After localizing PF joint ROI, a deep convolutional neural network (CNN) model was used for predicting

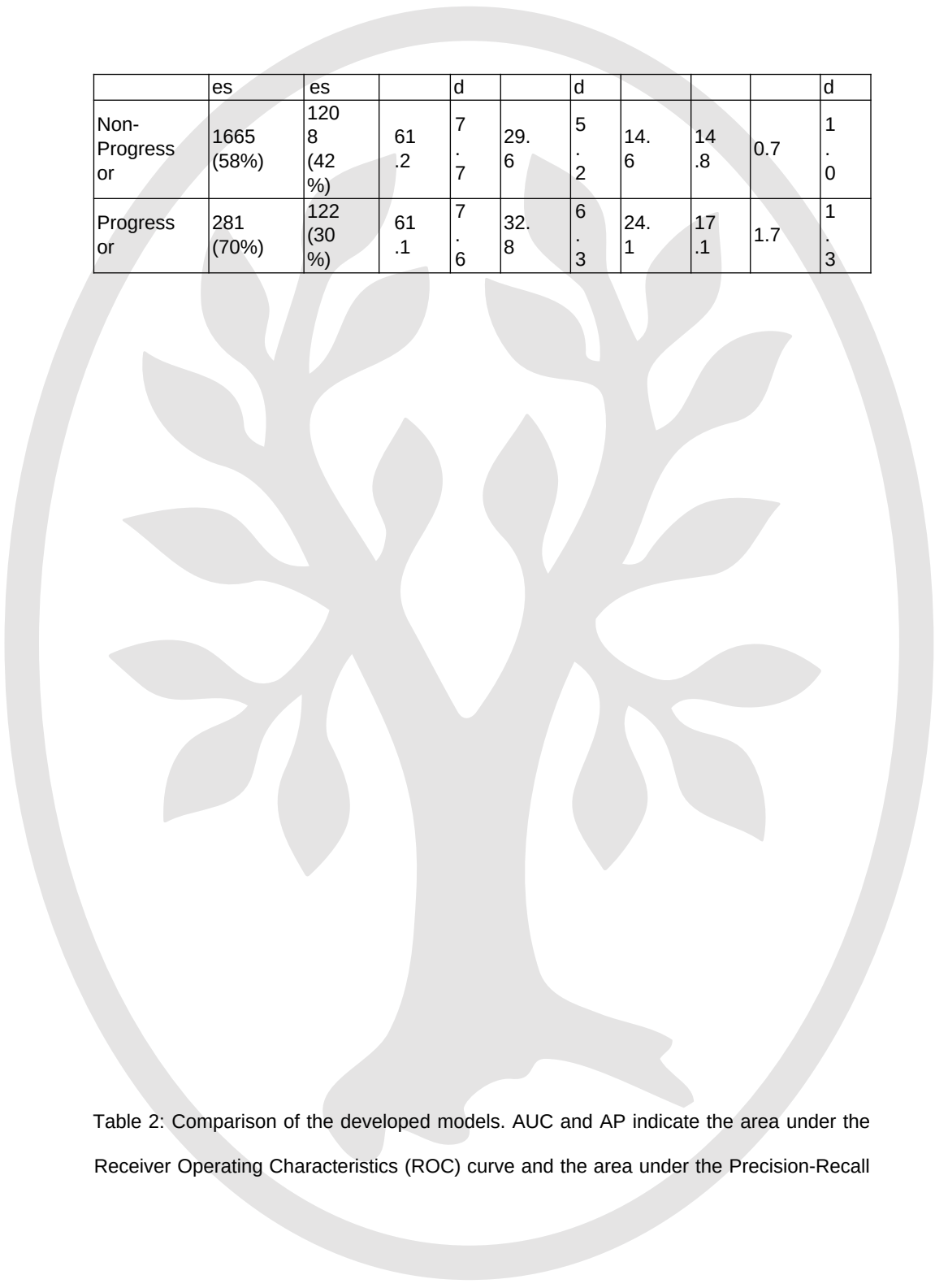
the progression of patellofemoral osteoarthritis (PFOA). b) For comparison, a separate machine learning model (gradient boosting machine (GBM)) was trained based on clinical variables including age, sex, body mass index (BMI), the total Western Ontario and McMaster Universities Arthritis Index (WOMAC) score, and Kellgren and Lawrence (KL) score of the tibiofemoral joint. We used a stratified subject-wise 5-fold cross validation setting to measure the performance of all the models. c) In addition to these individual models, we fused the predictions from these models in a second layer GBM model to improve the overall prediction performance

Figure 2: Chart shows the selection of knees from the MOST study used in this work.

Figure 3: Example of PFOA progression/development. Figure on the left demonstrates an exemplar patellofemoral joint ROI imaged at the first visit in the MOST study. At the baseline, PFOA is not present. Right figure presents the same participant's PF joint 7 years after the baseline visit. The knee has developed PFOA where joint space narrowing (JSN) and osteophytes - characteristic features of OA - are clearly seen. Best viewed on screen

Figure 4: Illustration of automated ROI localization. First, patellar height (h) was determined using landmarks. Subsequently, a small margin (Δ) is padded around the patellar region. On the femur side, ROI is located such that the width equals to the height of the ROI. Best viewed on screen.

Figure 5: (a) ROC and (b) PR curves demonstrating the performance of the models. Area under the curves and 95% confidence intervals in parentheses were given based on a 5-fold cross validation setting. Dashed lines in ROC indicate the performance of a random

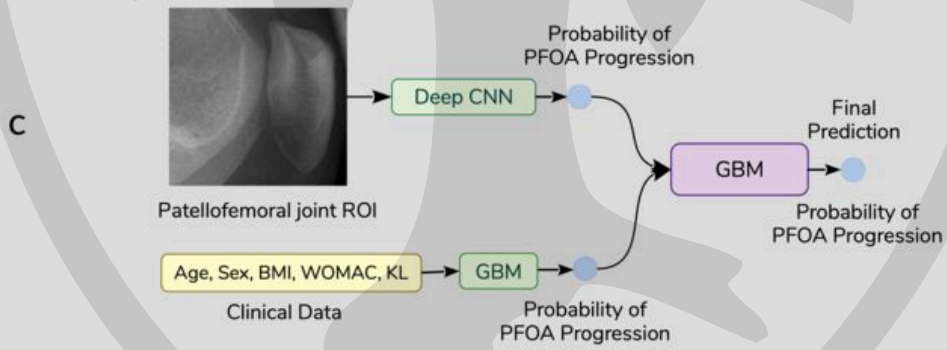
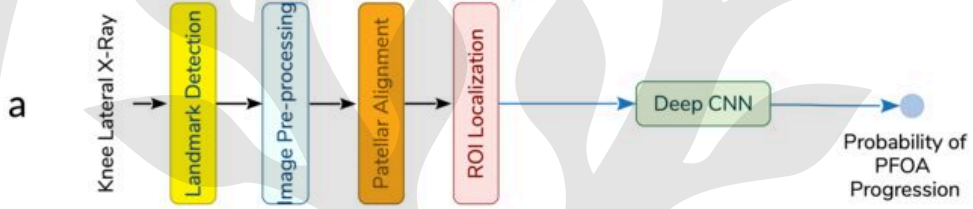
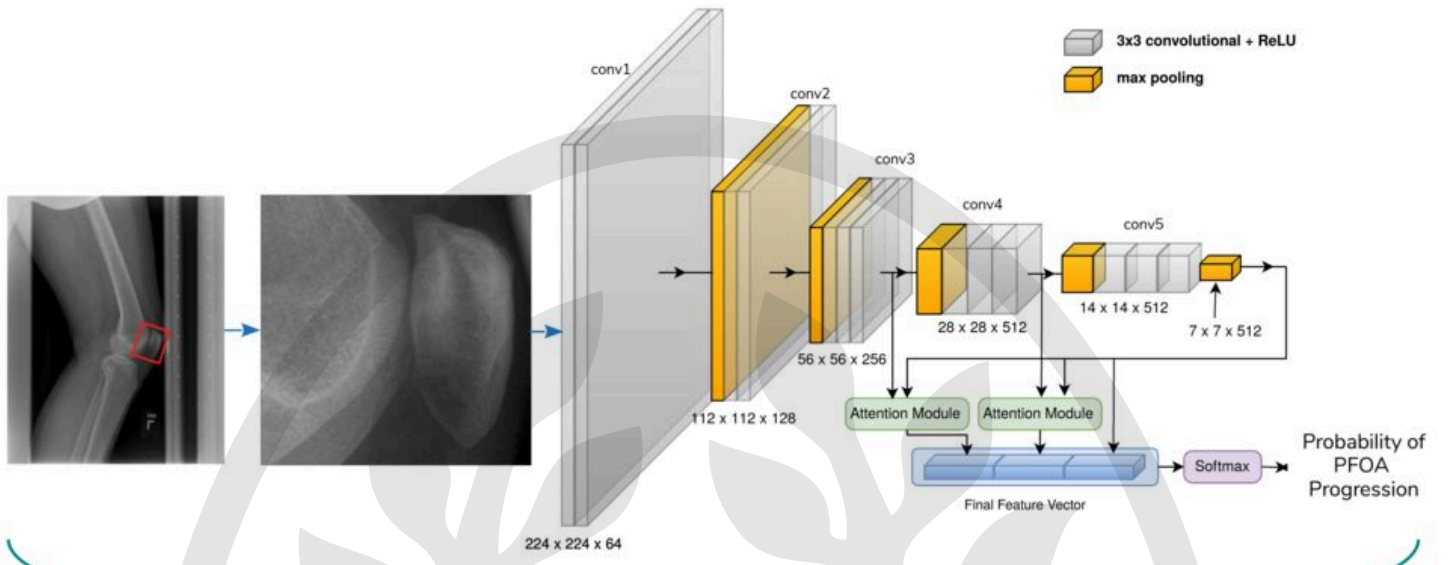


| | es | es | | d | | d | | | | d |
|----------------|---------------|---------------|------|-----|------|-----|------|------|-----|-----|
| Non-Progressor | 1665 (58%) | 1208 (42%) | 61.2 | 7.7 | 29.6 | 5.2 | 14.6 | 14.8 | 0.7 | 1.0 |
| Progressor | 281 (70%) | 122 (30%) | 61.1 | 7.6 | 32.8 | 6.3 | 24.1 | 17.1 | 1.7 | 1.3 |

Table 2: Comparison of the developed models. AUC and AP indicate the area under the Receiver Operating Characteristics (ROC) curve and the area under the Precision-Recall

(PR) curves, respectively. The 95% confidence intervals in parentheses were given based on a 5-fold cross validation setting.

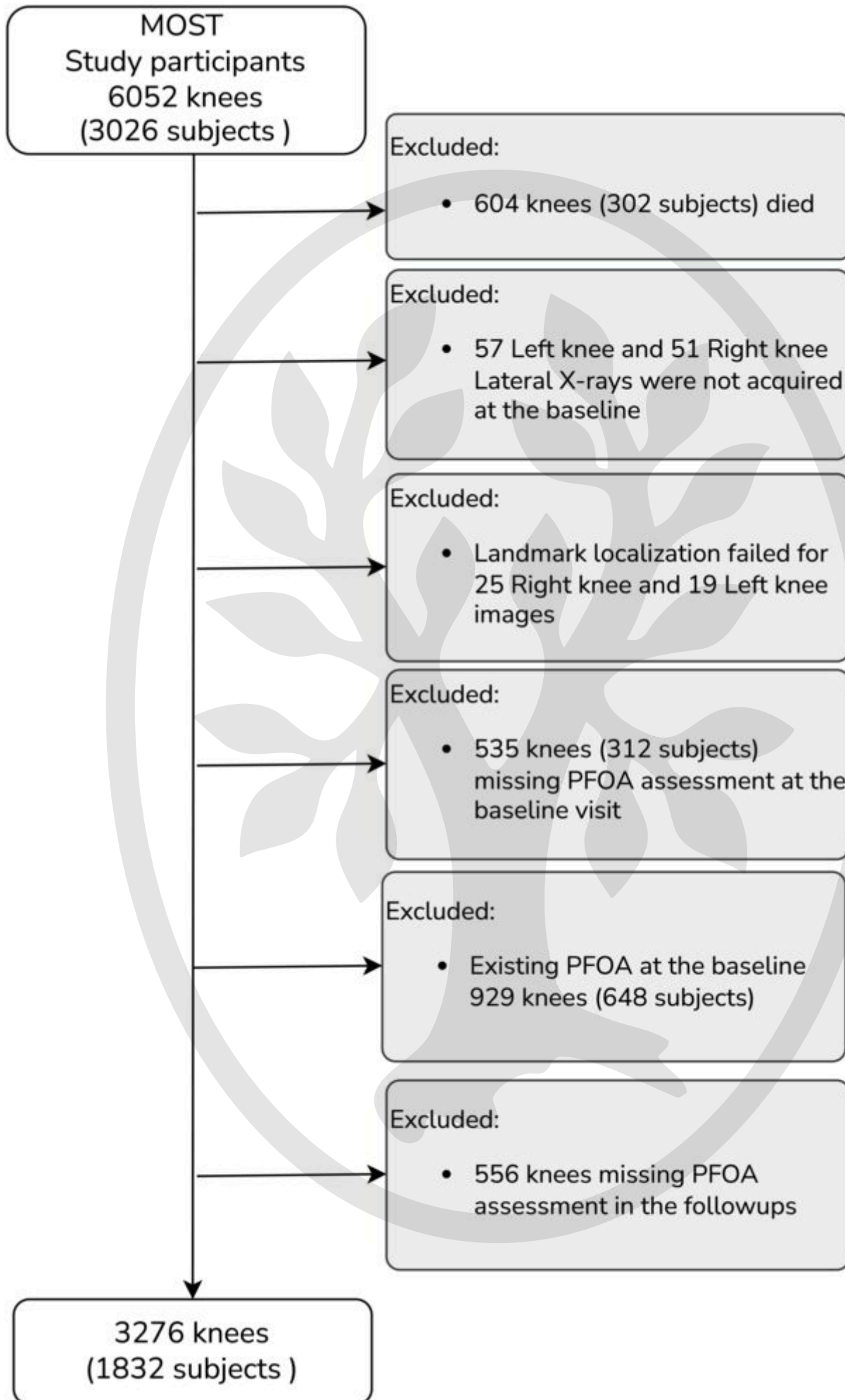
| | Input | Method | AUC [95% CI] | AP [95% CI] | Brier Score | |
|---------------|---------------------------------------|--------|--------------------------------|--------------------------------|----------------|------------------|
| <i>Model1</i> | Age, Sex, BMI | GBM | 0.655 | 0.232 | 0.103 | Clinical |
| <i>Model2</i> | Age, Sex, BMI, | GBM | 0.707 | 0.265 | 0.100 | Clinical |
| <i>Model3</i> | Age, Sex, BMI, | GBM | 0.767 [0.74, | 0.334 | 0.095 | Clinical |
| <i>Model4</i> | VGG-16 | CNN | 0.832 | 0.4 [0.359, | 0.262 | CNN |
| <i>Model5</i> | VGG-16-Attrn | CNN | 0.856 | 0.431 | 0.165 | CNN |
| <i>Model6</i> | Predictions from Model3 and Model5 | GBM | 0.865 [0.849, | 0.447 [0.404, | 0.084 | Stacked Model |

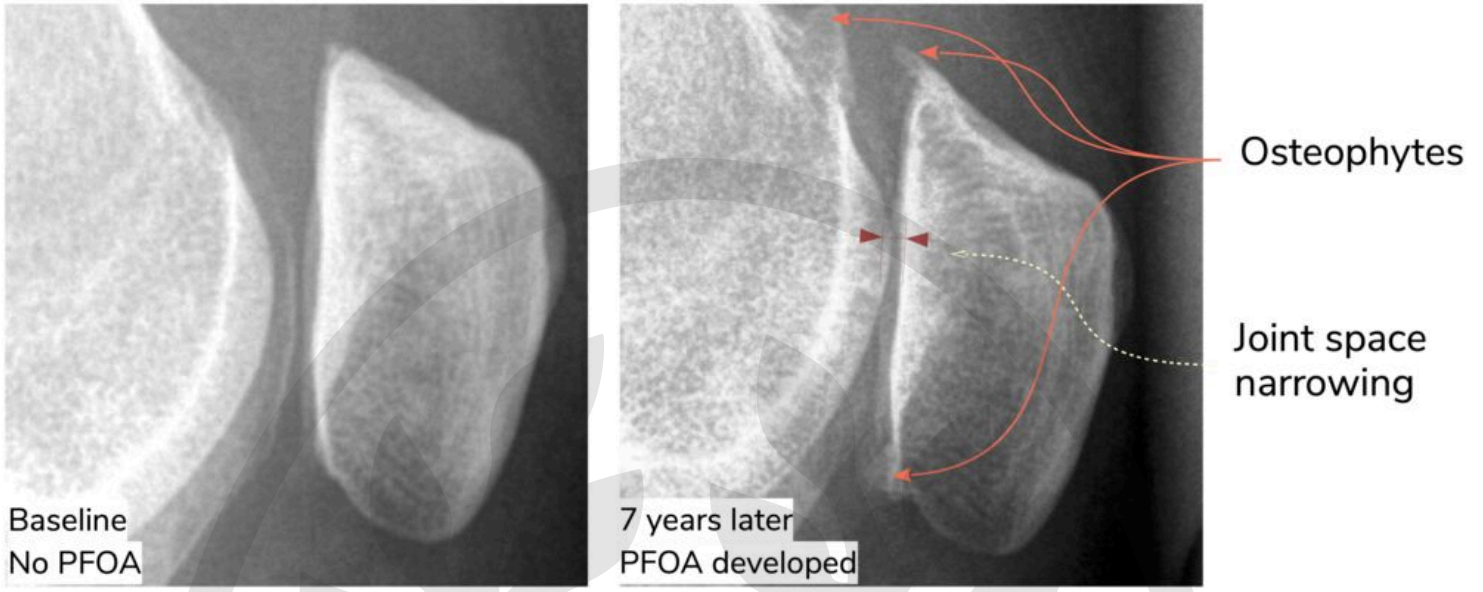


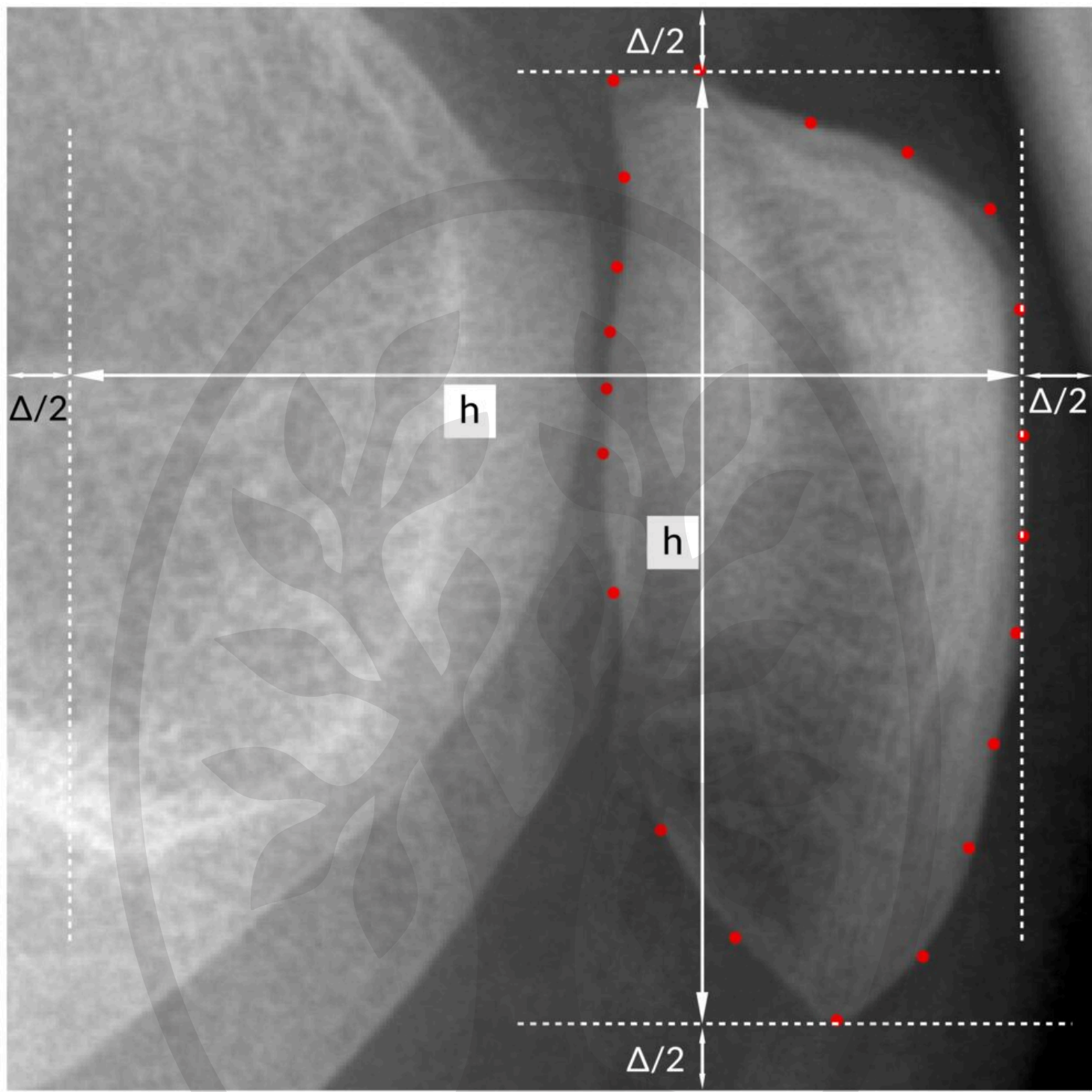
Deep CNN Model

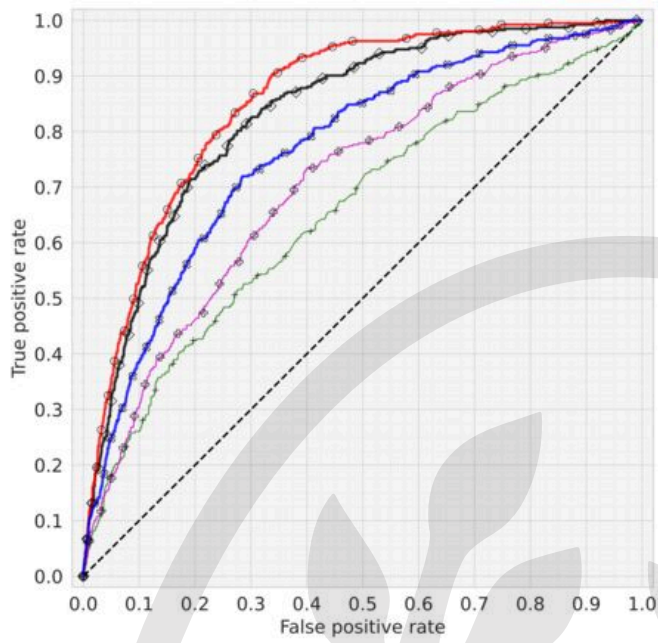
Reference Models

Stacked Model

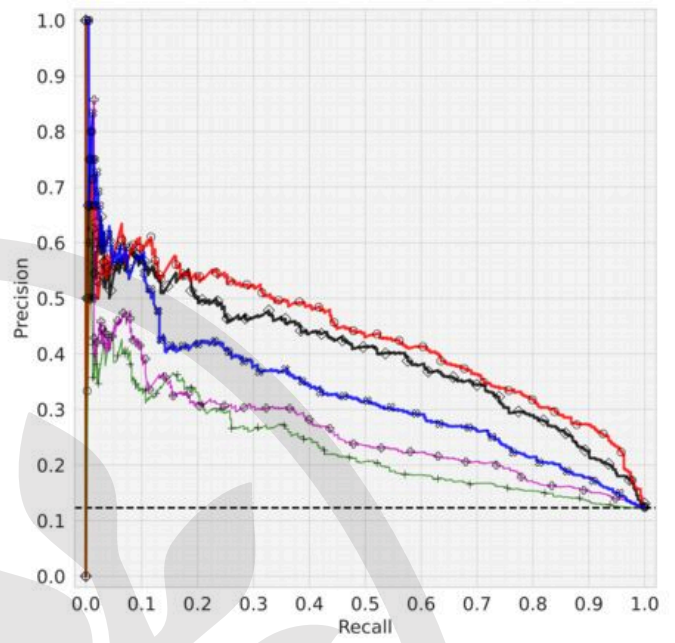




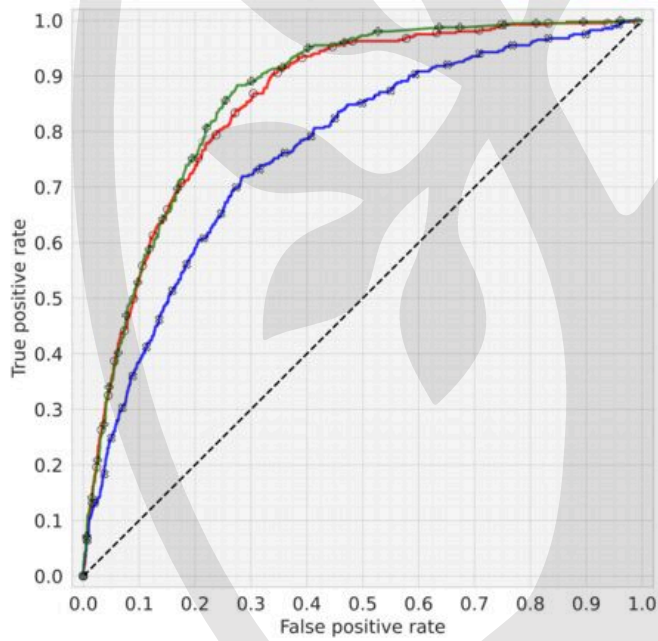




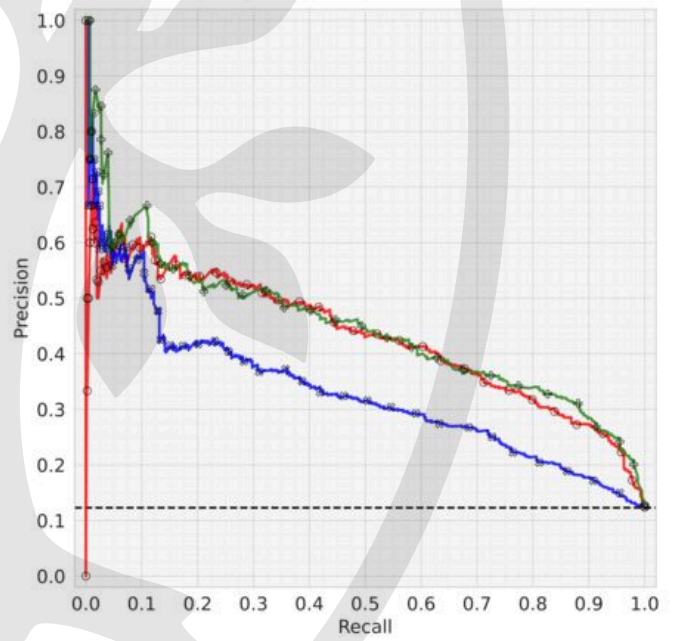
— Vgg (AUC= 0.832 [0.812, 0.851])
 — Vgg-Attn (AUC= 0.856 [0.838, 0.872])
 — AGE, SEX, BMI (AUC= 0.655 [0.624, 0.684])
 — AGE, SEX, BMI, WOMAC (AUC= 0.707 [0.678, 0.732])
 — AGE, SEX, BMI, WOMAC, KL (AUC= 0.767 [0.74, 0.789])



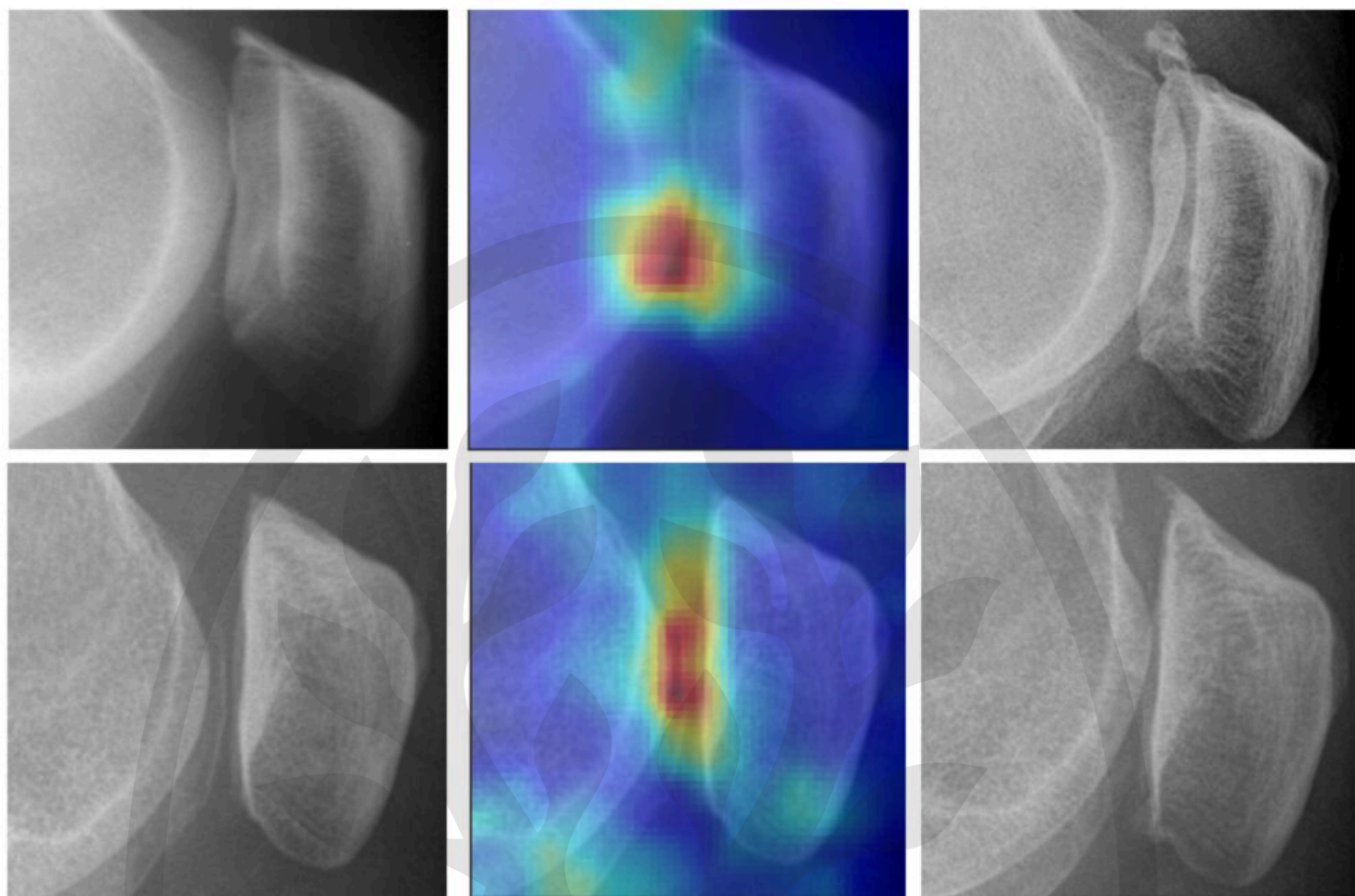
— Vgg (AP= 0.4 [0.359, 0.444])
 — Vgg-Attn (AP= 0.431 [0.387, 0.475])
 — AGE, SEX, BMI (AP= 0.232 [0.205, 0.268])
 — AGE, SEX, BMI, WOMAC (AP= 0.265 [0.231, 0.299])
 — AGE, SEX, BMI, WOMAC, KL (AP= 0.334 [0.293, 0.377])



— Vgg-Attn (AUC= 0.856 [0.838, 0.872])
 — AGE, SEX, BMI, WOMAC, KL (AUC= 0.767 [0.74, 0.789])
 — Vgg-Attn + AGE_SEX_BMI_WOMAC_KL (AUC= 0.865 [0.849, 0.88])



— Vgg-Attn (AP= 0.431 [0.387, 0.475])
 — AGE, SEX, BMI, WOMAC, KL (AP= 0.334 [0.293, 0.377])
 — Vgg-Attn + AGE_SEX_BMI_WOMAC_KL (AP= 0.447 [0.404, 0.491])



Baseline visit
No PFOA

Attention Map

7 years visit
PFOA developed