SCIENTIFIC REPORTS
nature research

OPEN

Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data

Aleksei Tiulpin[1,8*], Stefan Klein[2], Sita M. A. Bierma-Zeinstra[3,4], Jérôme Thevenot[1], Esa Rahtu[5], Joyce van Meurs[6], Edwin H. G. Oei[7] & Simo Saarakkala[1,8]

Knee osteoarthritis (OA) is the most common musculoskeletal disease without a cure, and current treatment options are limited to symptomatic relief. Prediction of OA progression is a very challenging and timely issue, and it could, if resolved, accelerate the disease modifying drug development and ultimately help to prevent millions of total joint replacement surgeries performed annually. Here, we present a multi-modal machine learning-based OA progression prediction model that utilises raw radiographic data, clinical examination results and previous medical history of the patient. We validated this approach on an independent test set of 3,918 knee images from 2,129 subjects. Our method yielded area under the ROC curve (AUC) of 0.79 (0.78–0.81) and Average Precision (AP) of 0.68 (0.66–0.70). In contrast, a reference approach, based on logistic regression, yielded AUC of 0.75 (0.74–0.77) and AP of 0.62 (0.60–0.64). The proposed method could significantly improve the subject selection process for OA drug-development trials and help the development of personalised therapeutic plans.

Knee osteoarthritis (OA) is the most common musculoskeletal disorder causing significant disability for patients worldwide[1]. OA is a degenerative disease and there is a lack of knowledge on the factors contributing to its progression. The overall etiology of OA is also not understood and there is no effective treatment, besides behavioral interventions. Furthermore, at the end stage of the disease, the only available treatment option is total knee replacement (TKR) surgery, which is highly invasive, costly and also strongly affects the patient's quality of life. OA is a major burden for the public health care system and it is increasing further with the aging of the population. For example, according to the statistics only in the United States, around 12% of the population suffer from OA and the annual rate of TKR for people 45–64 years of age has doubled since the year of 2000[2]. From the economical point of view, OA causes enormous costs for society and the costs of these surgeries are estimated to be over nine billion euros[2].

In primary health care, OA is currently diagnosed based on a combination of clinical history, physical examination, and X-ray imaging (radiography) if needed. However, the current widely available diagnostic modalities do not allow for effective OA prognosis assessment[3], which is important for the planning of appropriate therapeutic interventions and also for recruitment to OA disease modifying drugs development trials[4]. A possible improvement would be to extend this diagnostic chain with Magnetic Resonance Imaging (MRI), which is, however, costly, time-consuming, has limited availability and not applicable for wide use[5].

While being imperfect and lacking decision consistency, the current OA diagnostic tools can be enhanced using computer-assisted methods. For example, it has been shown that the gold clinical standard for OA severity assessment from radiographs, semi-quantitative Kellgren-Lawrence (KL)[6] system that highly suffers from

[1]Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland. [2]Biomedical Imaging Group Rotterdam, Depts. of Medical Informatics & Radiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [3]Department of General Practice, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [4]Department of Orthopedics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [5]Department of Signal Processing, Tampere University of Technology, Tampere, Finland. [6]Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [7]Department of Radiology & Nuclear Medicine, University Medical Center Rotterdam, Rotterdam, The Netherlands. [8]Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland. *email: aleksei.tiulpin@oulu.fi

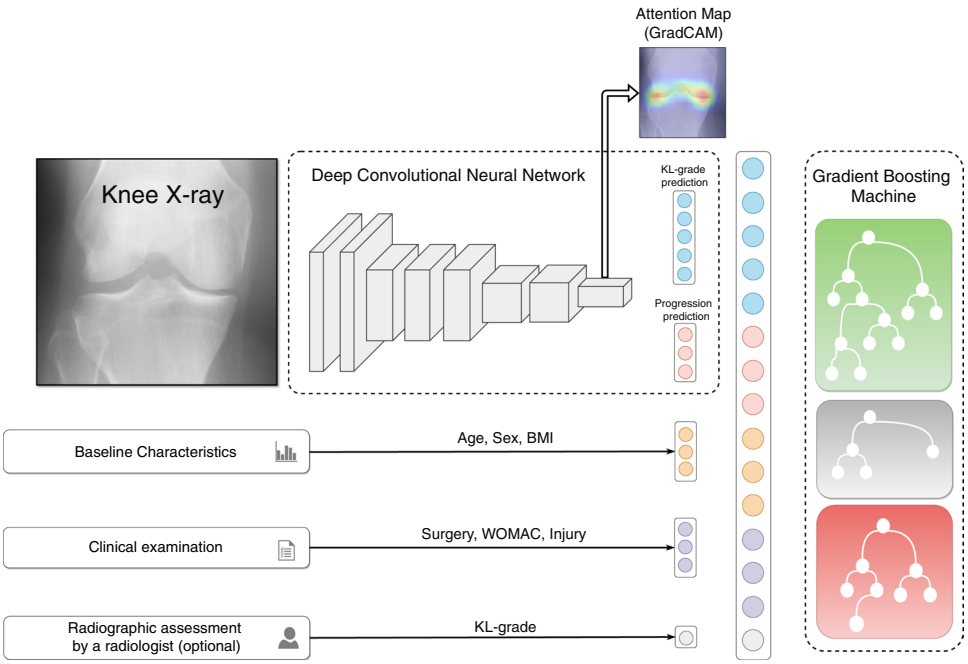

Figure 1. Schematic representation of our multi-modal pipeline, predicting the risk of osteoarthritis (OA) progression for a particular knee. We first use a Deep Convolutional Neural Network (CNN), trained in a multi-task setting to predict the probability of OA progression (no progression, rapid progression, slow progression) and the current stage of OA defined according to the Kellgren-Lawrence (KL) scale. Subsequently, we fuse these predictions with patient's Age, Sex, Body-Mass Index, given knee injury and surgery history, symptomatic assessment results and, optionally, a KL grade given by a radiologist using a Gradient Boosting Machine Classifier. After obtaining prediction from CNN, we utilize GradCAM attention maps to make our method more transparent and highlight the zones in the input knee radiograph, which were considered most important by the network.

subjectivity of a practitioner, can be automated using Deep Learning – a state-of-the-art Machine Learning approach widely used in computer vision[7–9]. However, to the best of our knowledge, there have been no similar studies on Deep Learning-based prediction of structural knee OA progression, in which the raw image data are directly used for prediction instead of the KL grades defined by a radiologist.

Current state-of-the-art OA progression prediction models are based on a combination of texture descriptors extracted from trabecular bone (e.g. fractal signature analysis)extracted from trabecular bone (e.g. fractal signature analysis), KL-grade, clinical and anthropometric data[10–13]. However, their performance and generalizability are difficult to assess for multiple reasons. Firstly, the texture descriptors may suffer from sensitivity to data acquisition settings. This can lead to limited sample size, as is for example seen in the studies of Janvier *et al.*, where only non-processed digital images were used[11,12]. Secondly, only a few current progression studies used an external dataset besides the one that was utilized to develop the prediction model[10,14,15]. If such external dataset is not utilised, this can lead to a possible overfitting and eventually a bias in the final results[7]. Finally, it has been previously shown that most of the OA evolution modelling studies tend to focus on estimating the decrease of joint space width (JSW) as a measure of progression[16]. Such outcome can be challenging to validate due to inherent problems associated with radiographic data acquisition (e.g. varying beam angle) and it does not depict all the changes happening with the joint. Previously, it has been recommended to assess OA progression using measures that incorporate the information about both – JSW and osteophytes[17], *i.e.*, treating future increase of the KL-grade as a progression outcome.

In this study, we propose a novel method based on Machine Learning that directly utilises raw radiographic data, physical examination, patient's medical history, anthropometric data and, optionally, a radiologist's statement (KL-grade) to predict structural OA progression. Here, we aim to predict any increase of a current KL-grade or potential need for TKR within the next 7 years after the baseline examination for patients having no, early or moderate OA. One of the main strengths of this study is that we used a separate large dataset for independent testing of our and the reference approachesOne of the main strengths of this study is that we used a separate large dataset for independent testing of our and the reference approaches. The proposed method employs a Deep Convolutional Neural Network (CNN)[18,19] that evaluates the probability of OA progression jointly with the current OA severity in the analysed knee as an auxiliary outcome. Further, we improve the prognosis from CNN by fusing its prediction with the clinical data using a Gradient Boosting Machine (GBM)[20]. Schematically, our method is presented in Fig. 1.

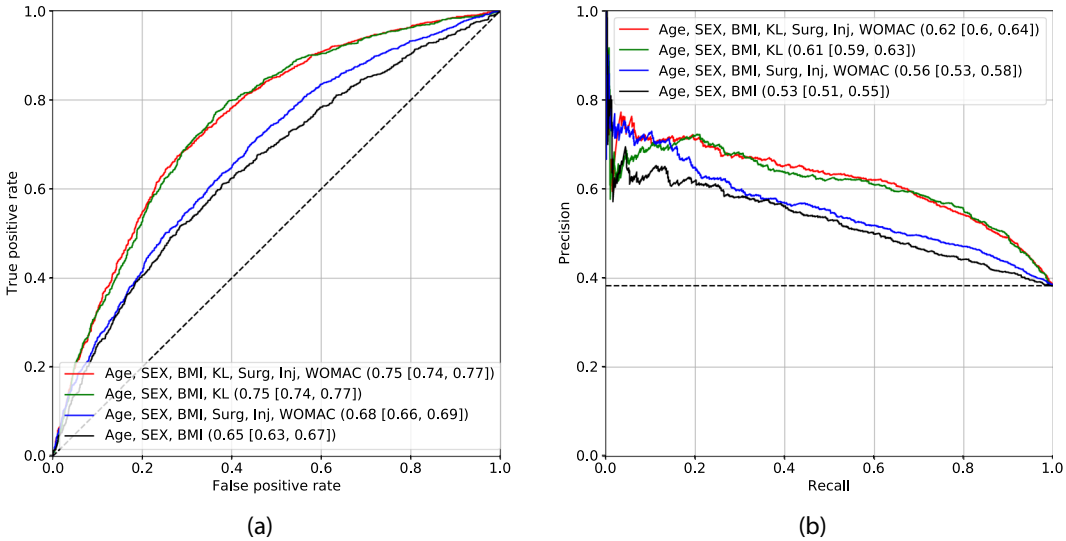

Figure 2. Assessment of Logistic Regression-based models' performance. The subplot (**a**) demonstrates the ROC curves and the subplot (**b**) precision-recall curves. Black dashed lines indicate the performance of a random classifier in case of AUC, and performance of the prediction model based on the dataset labels distribution. The subplots' legends reflect the benchmarked models and the values of corresponding metrics with 95% confidence intervals. Here, Area under the ROC curve metric is used in subplot (**a**) and Average Precision in subplot (**b**).

Results

Training and testing datasets. We used the metadata provided in Osteoarthritis Initiative (OAI) and Multicenter Osteoarthritis Study (MOST) cohorts to select progressors and non-progressors for train and test datasets, respectively. We considered only the knees having no, early or moderate OA (KL-0, KL-1, KL-2 and KL-3) at the baseline (first visit) as these are the most relevant clinical cases. Furthermore, we excluded from the test set all the subjects who died between the follow-ups for coherence of our data. Additionally, the subjects who did not progress and dropped out from the study before the last follow-up examination were excluded. After the pre-selection process, we used 4,928 knees (2,711 subjects) from OAI dataset for training and 3,918 knees (2,129 subjects) from MOST dataset for testing of our model. Here, 1,331 (27%) and 1,501 (47%) knees were identified as progressors in OAI and MOST data, respectively. As a progression definition, we utilised an increase of a KL-grade within the following years. Here, we ignored the increase from KL-0 to KL-1 and included all cases with progression to TKR. To harmonise the data between OAI and MOST datasets, we defined the following three fine-grained categories:

- $y = 0$: no knee OA progression
- $y = 1$: progression within the next 60 months (fast progression)
- $y = 2$: progression after 60 months (slow progression)

Supplementary Tables S2 and S3 describe the training and the testing sets derived from OAI and MOST datasets respectively.

Reference methods. Firstly, we utilised several reference methods (see details in Methods) in order to understand the added value of our approach. These models were trained to predict a probability $P(y > 0|x)$ of a particular knee x to have a KL-grade increase in the future. Here, we pooled the classes $y = 1$ and $y = 2$ together to derive a binary outcome, which was used in both Logistic Regression (LR) and GBM reference methods. In Fig. 2, we demonstrate the performance of LR, which is commonly used in OA research[10,11,14,15]. All of the LR models were derived and tested on the existing image assessment and clinical data provided by the OAI and MOST datasets, respectively. In cross-validation experiments on OAI data, we also assessed the added value of regularization[21] and found no difference between regularised and non-regularised LR models.

From Fig. 2, it can be seen that two best models exist: one based on Age, Sex, Body-Mass Index and KL grade (model 1), and the other being the same with the addition of symptomatic assessment (Western Ontario and McMaster Universities Arthritis Index, WOMAC[22]), injury and surgery history (model 2). We chose the latter in our further comparisons because it performs with higher precision at lower recall while yielding similar performance at other recall levels. This model yielded AUC of 0.75 (0.74–0.77) and Average Precision (AP) of 0.62 (0.60–0.64). All the mentioned risk factors included into the reference models were selected on the basis of their use in the previous studies[10,14,15].

It was hypothesised that LR might not be able to exploit the full potential of the input data (clinical variables and image assessments), as with this type of model, non-linear relationships within the data cannot be evaluated. Therefore, we utilised a GBM and trained it to predict the probability of OA progression. Figure 3 demonstrates

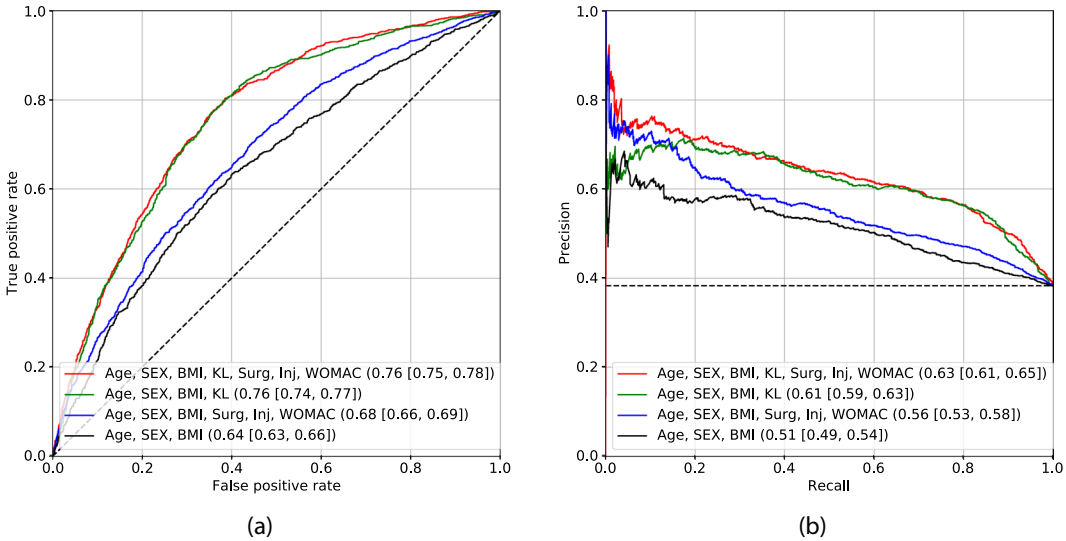

Figure 3. Assessment of Gradient Boosting Machine-based models' performance. The subplot (**a**) demonstrates the ROC curves and the subplot (**b**) precision-recall curves. Black dashed lines indicate the performance of a random classifier in case of AUC, and performance of the prediction model based on the dataset labels distribution. The subplots' legends reflect the benchmarked models and the values of corresponding metrics with 95% confidence intervals. Here, Area under the ROC curve metric is used in subplot (**a**) and Average Precision in subplot (**b**).

Model	AUC		AP	
	LR	GBM	LR	GBM
Age, Sex, BMI	0.65 (0.63–0.67)	0.64 (0.63–0.66)	0.53 (0.51–0.55)	0.52 (0.49–0.54)
Age, Sex, BMI, Injury, Surgery, WOMAC	0.68 (0.66–0.69)	0.68 (0.66–0.69)	0.56 (0.53–0.58)	0.56 (0.53–0.58)
KL-grade	0.73 (0.71–0.75)	—	0.57 (0.55–0.58)	—
Age, Sex, BMI, KL-grade	0.75 (0.74–0.77)	0.76 (0.74–0.77)	0.61 (0.59–0.63)	0.61 (0.59–0.63)
Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade	0.75 (0.74, 0.77)	0.76 (0.75–0.78)	0.62 (0.60–0.64)	0.63 (0.61–0.65)

Table 1. Summary of the reference models' performances on the test set. Top performing models are underlined. 95% confidence intervals are reported in parentheses. BMI – Body-Mass Index. WOMAC – Western Ontario and McMaster Universities Arthritis Index. KL-grade – Kellgren-Lawrence grade. AUC – Area Under the Receiver Operating Characteristic Curve. AP – Average Precision. LR – Logistic Regression. GBM – Gradient Boosting Machine.

the performance of models identical to model 1 and model 2, but trained using GBM instead of LR (model 3 and model 4). Model 4 performed best and obtained the AUC of 0.76 (0.75–0.78) and AP of 0.63 (0.61–0.65). The full comparisons of the models built using LR and GBM approaches are summarised in Table 1 and also in Figs. 2 and 3.

Predicting progression from raw image data. After testing the reference models, we developed a CNN, which allows to directly leverage raw knee DICOM images in an automatic manner. In contrast to the previous studies, this model was trained in a multi-task setting to predict OA progression in the index knee and also its current KL-grade from the corresponding X-ray image. In particular, our model consists of a feature extractor – a pre-trained se-resnext50-32xd model[23] – and two branches, each of which is a fully connected layer (FC), predicting its own task. One branch of the model predicts a progression outcome and the other branch a KL grade (Fig. 1).

In our experiments, we found that prediction of the previously defined fine-grained classes – no ($y = 0$), fast ($y = 1$) and slow ($y = 2$) progression, while being inaccurate individually, helps to regularize the training of the CNN and leads to better performance in predicting overall probability of progression $P(y > 0|x)$ within the following years. Having predicted such binary outcome, our CNN model (model 5) trained using the baseline knee image yielded AUC of 0.76 and AP of 0.56 in a cross-validation experiment on the training set. On the test set, the CNN yielded AUC of 0.79 (0.77–0.80) and AP of 0.68 (0.66–0.70). We compared this model to the strongest reference method – model 4, and also strongest conventional method based on LR – model 2 (Fig. 4). We obtained a statistically significant performance difference in AUC (DeLong's p-value $< 1e-5$) when compared our CNN to the model 4.

To gain insight into the basis of the CNN's prediction, we used the GradCAM[24] approach and visualised the attention maps for the well-predicted knees. Examples of attention maps are presented in Fig. 5. We observed

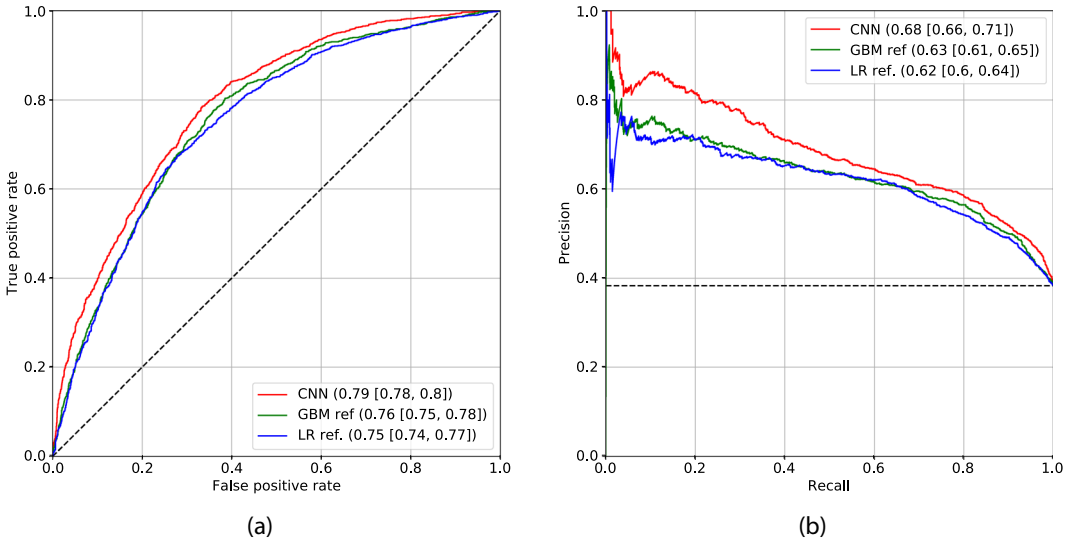

Figure 4. Comparison of the deep convolutional neural network (CNN) and the reference methods built using Gradient Boosting Machine (GBM). Reference method based on Logistic Regression is also presented for better visual comparison (model 2 in the text). CNN model utilises solely knee image and the GBM model utilises KL grade and clinical data (model 4 in the text). Subplot (**a**) shows the ROC curves for CNN and GBM respectively. Subplot (**b**) shows the Precision-Recall Curves. Black dashed lines indicate the performance of a random classifier in case of AUC, and performance of the prediction model based on the dataset labels distribution. The subplots' legends reflect the benchmarked models and the values of corresponding metrics with 95% confidence intervals. Here, Area under the ROC curve metric is used in subplot (**a**) and Average Precision in subplot (**b**).

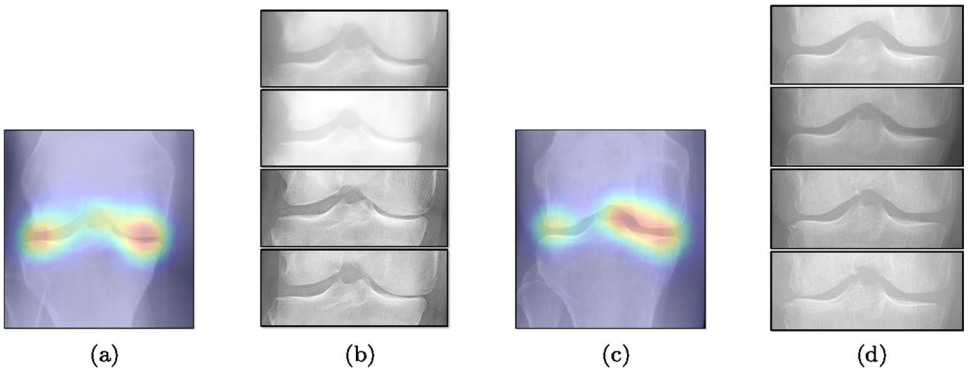

Figure 5. Examples of attention maps for progression cases and the corresponding visualization of progression derived using follow-up images from MOST datasets. Here, subplots (**a**,**c**) show the attention maps derived using a GradCAM approach. Subplots (**b**,**d**) show the joint-space areas from all the follow-up images (baseline to 84 months). Here, the subplot (**b**) corresponds to the attention map (**a**) and the subplot (**d**) corresponds to the attention map (**c**).

that in various cases, the CNN paid attention to the compartment opposite to the one where degenerative change became visible during the follow-up visits. Additional examples of such attention maps are presented in Supplementary Figs. S6, S7, S8 and S9.

To evaluate whether a combination of conventional diagnostic measures used in models 1–4 and CNN would further increase the predictive accuracy, we utilised a GBM in a stacked generalisation fashion[25] and treated both clinical measures and CNN's predictions as input features for the GBM (see Fig. 1). Two stacked models were created. The first model, model 6, is fully automatic (does not use a KL-grade as an input) and predicts a probability of OA progression. It was built using all the predictions produced by the CNN – $P(KL = i|x)$ for $i \in \{0, \ldots, 3\}$ and $P(y = i|x)$ for $i \in \{0, \ldots, 2\}$, and additionally age, sex, BMI, knee injury history, knee surgery history and WOMAC total score. The second model, model 7, was similar to the model 6, but with the addition of the KL grade that provides additional source information about the current stage of OA to the GBM. More details on building and training this two-stage pipeline are given in Methods. We hypothesised that a radiologist and a neural network may assign a KL grade differently, therefore, the difference in gradings could be leveraged for the prediction model, *e.g.* if these gradings differ.

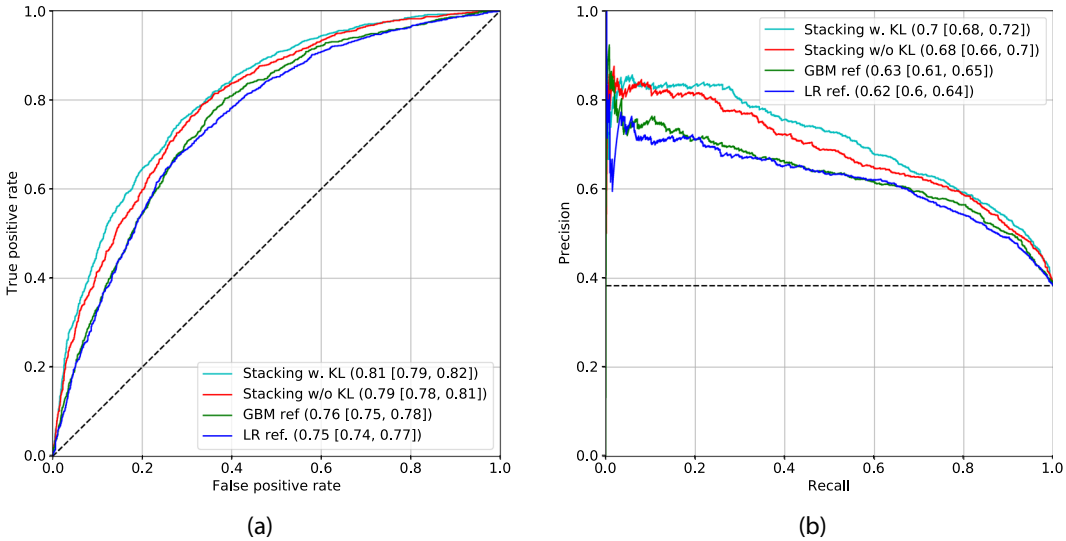

Figure 6. Comparison of the multi-modal methods, based on Deep Convolutional Neural Network (CNN) and Gradient Boosting Machine (GBM) classifier versus the strongest reference method (model 4). Reference method based on Logistic Regression is also presented for better visual comparison (model 2). The subplots' legends reflect the benchmarked models and the values of corresponding metrics with confidence intervals. Black dashed lines indicate the performance of a random classifier in case of AUC, and performance of the prediction model based on the dataset labels distribution. Here, Area under the ROC curve is used in subplot (**a**) and Average Precision in subplot (**b**). The subplots (**a**,**b**) show the ROC and Precision-Recall (PR) curves respectively. The results in this plot indicate that our method benefits from the utilization of a KL-grade.

Model #	Model	AUC	AP
2	Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade (LR)	0.75 (0.74–0.77)	0.62 (0.60–0.64)
4	Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade (GBM)	0.76 (0.75–0.78)	0.63 (0.61–0.65)
5	CNN	0.79 (0.77–0.80)	0.68 (0.66–0.70)
6	CNN + Age, Sex, BMI, Injury, Surgery, WOMAC (GBM-based fusion)	0.79 (0.78–0.81)	0.68 (0.66–0.71)
7	CNN + Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade (GBM-based fusion)	0.81 (0.79–0.82)	0.70 (0.68–0.72)

Table 2. Detailed comparison of the developed models for all subjects included into testing conducted on the MOST dataset. 95% confidence intervals are reported in parentheses for each of the reported metric. KL-grade – Kellgren-Lawrence grade. CNN – Deep Convolutional Neural Network. BMI – Body-Mass Index. WOMAC – Western Ontario and McMaster Universities Arthritis Index. AUC – Area Under the Receiver Operating Characteristic Curve. AP – Average Precision. LR – Logistic regression. GBM – Gradient Boosting Machine.

Figure 6 shows the ROC and PR curves of models 6 and 7, along with the best reference method, model 4. As reported earlier, this reference model yielded AUC of 0.76 (0.75–0.78) and AP of 0.63 (0.61–0.65). In contrast, our multi-modal methods without and with utilization of a KL grade – model 6 and model 7, yielded AUC of 0.79 (0.78–0.81), AP of 0.68 (0.66–0.71) and AUC of 0.80 (0.79–0.82), AP of 0.70 (0.68–0.72) respectively. Additionally, we also show the ROC and PR curves for model 2 in Fig. 6. In Table 2, we present a detailed comparison of models 2, 4, 5, 6 and 7. In addition to the performance metrics, we conducted analyses of feature importance for both train and test datasets (Supplementary Figs. S4 and S5, respectively). These analyses showed a strong impact of the CNN's predictions onto the output of the GBM for both, models 6 and 7, respectively. Besides, in the case of model 7, we observed a strong impact of a KL-grade on the final prediction. Further discussion and description of the analyses are presented in Supplementary Experiments. In addition to the performance metrics, we conducted analyses of feature importance for both train and test datasets (Supplementary Figs. S4 and S5, respectively). These analyses showed a strong impact of the CNN's predictions onto the output of the GBM for both, models 6 and 7, respectively. Besides, in the case of model 7, we observed a strong impact of a KL-grade on the final prediction. Further discussion and description of the analyses are presented in Supplementary Experiments.

Finally, we also present the results on predicting OA progression for the subgroup of knees identified as KL-0 or KL-1 at baseline. These results are presented in Table 3. The results for this particular group of knees show that our method is capable of identifying knees that will progress to OA in a fully automatic manner with high performance – our two best models, model 6 and model 7, yielded AUC of 0.78 (0.76–0.80) and 0.80 (0.78–0.82) respectively, and AP of 0.58 (0.55–0.62) and 0.62 (0.58–0.65) respectively.

Model #	Model	AUC	AP
2	Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade (LR)	0.73 (0.70–0.75)	0.52 (0.49–0.55)
4	Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade (GBM)	0.75 (0.72–0.77)	0.54 (0.51–0.58)
5	CNN	0.78 (0.76–0.80)	0.58 (0.55–0.61)
6	CNN + Age, Sex, BMI, Injury, Surgery, WOMAC (GBM-based fusion)	0.78 (0.76–0.80)	0.58 (0.55–0.62)
7	CNN + Age, Sex, BMI, Injury, Surgery, WOMAC, KL-grade (GBM-based fusion)	0.80 (0.78–0.82)	0.62 (0.58–0.65)

Table 3. Detailed comparison of the developed models for knees identified with Kellgren-Lawrence grade 0 or 1, which is considered as absence of osteoarthritis. The testing was done on the Multicenter Osteoarthritis Study dataset. 95% confidence intervals are reported in parentheses for each of the reported metric. KL-grade – Kellgren-Lawrence grade. CNN – Deep Convolutional Neural Network. BMI – Body-Mass Index. WOMAC – Western Ontario and McMaster Universities Arthritis Index. AUC – Area Under the Receiver Operating Characteristic Curve. AP – Average Precision. LR – Logistic regression. GBM – Gradient Boosting Machine.

Discussion

In this study, we presented a patient-specific machine learning-based method to predict structural knee OA progression from patient data acquired at a single clinical visit. The key difference of our method to the prior work is that it leverages the raw image of the patient's knee instead of any measures derived by human observers (e.g. JSW, KL or bonebone texture descriptors).

The results presented in this study demonstrate that our method yields significantly better prediction performance than the conventionally used reference methods. The major finding of this study is that it is possible to predict knee OA progression from a single knee radiograph complemented with clinical data in a fully automatic manner. Other findings of this study demonstrate that the knee X-ray image alone is already a very powerful source of data to predict whether a particular knee will have OA progression or not. Finally, one of the main results from a clinical point of view is that it is possible to predict progression for patients having KL-0 and KL-1 at baseline.

To the best of our knowledge, this is the first study where CNNs were utilised to predict OA progression directly from radiographs, and it is also one of the few studies in the field where an independent test set is used to robustly assess the results[10,14,15]. We believe that having such settings, where the test set remains unused until the final model's validation, is crucial for further development of the OA progression prediction models. Another novelty of our approach is leveraging multi-modal patient data: plain radiographs (raw image data compared to KL-grades used previously[10,15] or manually designed texture parameters[11,12]), symptomatic assessment, and patient's injury and/or surgery history data for prediction. Our results highlight that a combination of all the data allows to make more accurate predictions. Furthermore, thanks to GBM, with this approach it was possible to use missing data without imputation.

In principle, clinical application of the developed method is straightforward and makes it possible to detect OA progression at a low cost in primary health care with minimal modifications to the current diagnostic chain. Our method can be utilized in a fully-automatic manner without a radiologist's statement, and therefore, it could become available as an e.g. cloud service or software for physiotherapists to design behavioral interventions for the cases having high confidence of prediction. Compared to the other imaging modalities, such as MRI, the progression prediction methods developed just using radiographs and other easily obtainable data utilized in our study have potential to be the most accessible worldwide.

While machine learning-based approaches yield stronger prediction than conventional statistical models, (*e.g.* LR), they are less transparent, which can lead to lack of trust from clinicians. To address this drawback, various methods have been developed to explain the decisions of "black-box systems"[24,26,27]. As such, we utilised the GradCAM approach[24] that allowed us generating an attention map for each image at test timefor each image at test time, in order to highlight the zones where the CNN has paid its attention. While being attractive, this approach can also lead to wrong interpretations, i.e. there is no theoretical guarantee that the neural network identifies causal relationships between image features and the output variable. Therefore, a thorough analysis of the attention maps is required to assess the significance of certain features and anatomical zones picked-up by the model. Such analysis, however, could enable new possibilities for investigation of the visual features. For example, we observed interesting associations in the GradCAM-generated attention maps (Fig. 5), some of which are not captured by KL grading. As such, tibial spines (previously associated with OA progression[28]) were highlighted in multiple attention maps. These associations, however, do not hold for all the progressors.

From the attention maps, it can be seen that our model is hypothetically leveraging the information on JSW of the knee. We conducted multiple experiments solely on OAI dataset and verified whether our approach outperforms all our reference models that also include explicit measurements of JSW at fixed locations (fJSW)[29]. We found that model 7 outperforms any GBM-based model that includes fJSW measurements. These results are presented in Supplementary Table S1 and the detailed information regarding this result is presented in Supplementary Experiments.

Although our study demonstrates a novel method, which outperforms various state-of-the art reference approaches, it also has several important limitations. Firstly, our model has not been tested in other populations than the ones from the United States. Testing the developed model on data from other populations would be a crucial step to bring the developed machine learning-based approach to primary healthcare. Secondly, we utilised only standardised radiographs acquired with a positioning frame, which is not used in all the hospitals worldwide. Therefore, a validation of our model using the images acquired without the positioning frame is still

needed. However, we tried to address this limitation by including data acquired under different beam angles to the test set. Thirdly, we relied only on the KL-grading system to define a progression outcome, and the symptomatic component of OA progression was completely ignored. This also needs to be addressed in the future studies. Fourthly, we used imputation in the test set when evaluating LR models. This could potentially lower the performance of LR-based reference methods. In contrast, GBM-based approach allowed us to leverage all the samples with missing data without imputation. Fifthly, we would also like to mention the fact that we used total WOMAC score as a representation of patients' symptoms. While we think that this variable correlates with a symptoms check done in primary care, we also think that utilizing individual WOMAC components as features for GBM could potentially lead to a better performance of our method. Finally, our method is limited in terms of the requirements for training data. As such, in Supplementary Fig. S3 we demonstrated that the performance of our proposed CNN increases with the increased size of the train dataset (see Supplementary experiments for details). Consequently, future studies should consider enlarging the train dataset or improve the training techniques or CNN architecture.

The results presented in this study show that, for subjects at risk, our proposed knee OA progression prediction model allows to identify the progressor cases on average 6% more accurately than with the methods previously used in the OA literature. This study is an important step towards speeding up the OA disease modifying drug development process and also towards the development of better personalised treatment plans.

Methods

Data description and pre-processing. We utilised OAI (https://data-archive.nimh.nih.gov/oai) and MOST (http://most.ucsf.edu)OAI (https://data-archive.nimh.nih.gov/oai) and MOST (http://most.ucsf.edu) follow-up cohorts. Both of theseof these datasets include clinical and imaging data from subjects at risk of developing OA 45–79 and 50–79 years old, from baseline to 96 (9 imaging follow-ups) and 84 months (4 imaging follow-ups), respectively. OAI dataset includes bilateral posterior-anterior knee images, acquired with a Synaflexer™frame[30] and 10 degrees beam angle, while the MOST dataset also has images acquired with 5- and 15-degrees beam angles. OAI and MOST studies were approved by the institutional review board of the University of California San Francisco and also the data acquisition sites. The informed consent was obtained from all the subjects and all the data in both dataset is appropriately anonymised. All the protocols are available on the aforementioned web-sites for each of the cohorts. All the experiments with OAI and MOST datasets were performed in accordance with relevant guidelines and regulations.

Our inclusion criteria were the following. Firstly, we excluded the knees that had TKA, end-stage OA (KL-4) or had a missing KL-data at the baseline. Subsequently, we excluded the knees which did not progress and were not examined at the last follow-up. This allowed us to ensure that the subjects in the train and test sets did not progress within 96 and 84 months, respectively. If the knee had any increase of the KL-grade during the follow-up, we assigned the class of the earliest noticed KL-grade increase, e.g. if the knee progressed at 30 months and 84 months, we used 30-months follow-up visit to define the fine-grained progression class. Data selection flowcharts for OAI and MOST datasets are presented in Supplementary Figs. S1 and S2, respectively. The exact implementation of this selection process is also presented in the supplied source code (see Data Availability Statement).

In our experiments, we utilized variables such as age, sex, BMI, injury history, surgery history and total WOMAC (Western Ontario and McMaster Universities Arthritis Index) score. Due to the presence of missing values, it would be impossible to train and test LR model without utilizing imputation techniques or removing the missing data. Therefore, during the training of LR, we excluded the knees with missing values. In the test dataset (MOST), we imputed the missing variables by utilizing mean value imputation strategy when testing the LR. When we trained GBM-based method, the imputation strategies are not needed, thus we used the data extracted from OAI metadata as is.

Image pre-processing. To pre-process the OAI and MOST DICOM images, for each knee we extracted a region of interest (ROI) of 140×140 mm using an ad-hoc script and BoneFinder software[31] that enables accurate fully-automatic anatomicalfully-automatic anatomical landmark localization using regression voting approach. This was done in order to standardise the coordinate frame among the patients and the data acquisition centers. After localizing the bone landmarks, we rotated all the knee images so that the tibial plateau was horizontal. Subsequently, we performed a histogram clipping between 5^{th} and 99^{th} percentiles and used global contrast normalisation subtracting the image minimum and dividing all the image pixels by the maximum pixel value. Then, we converted the images to 8-bit depth multiplying them by 255. Finally, all the images were resised to 310×310 pixels (new pixel spacing of 0.45 mm) and the left knee images were flipped horizontally to match the collateral (right) knee.

In our initial experiments we tried to use 16-bit data and this had no effect on performance, but rather increased the size of the stored data. These experiments also included testing of different target pixel spacing, however, we eventually found that 0.45 mm spacing yielded the best results on cross-validation.

Experimental setup and reference methods. All experiments, including the hyper-parameter search, were carried out using the same 5-fold subject-wise cross-validation on OAI data. A stratified cross-validation was used to obtain the same distribution of progressed and non-progressed cases in both train and validation splits for each fold. To implement this validation scheme, we used the publicly available scikit-learn package[32].

For building regularised LR models, we used scikit-learn and for non-regularised LR we used the statsmodels package[33]. For GBM models, we utilised the LightGBM[34] implementation. We built the CNN models using PyTorch 1.0[35] and trained them using three NVidia GTX 1080Ti cards.

To find the best hyperparameters set for GBM, we used the Bayesian hyperparameters optimisation package hyperopt[36] with 500 trials. Each trial maximised the AP on cross-validation. In the case of CNN, we also used

cross-validation and built 5 models. We used the snapshot of the model's weights that yielded the maximum AP value on the validation set in each cross-validation split. The hyperparameters for CNN were found empirically.

Deep neural network's implementation details. We designed a multi-task CNN architecture to predict OA progression, and our model consisted of a convolutional (Conv) and two fully-connected (FC) blocks. One FC layer had three outputs corresponding to the three progression classes, and the other had 5 outputs, corresponding to the prediction of the current – baseline KL grade. This is schematically illustrated in Fig. 1. To harmonize the size of the outputs after Conv layers and the inputs of the FC layers, we utilised a Global Average Pooling layer.

We used the design of the Conv layers from se-resnext50_32x4d network[23]. In the initial cross-validation experiments, we also evaluated se-resnet50, inceptionv4, se-resnext101_32x4d; however, we did not obtain significantly better results than the ones reported in this study. To train the CNN, we utilised a transfer learning similarly to[7] and initialised the weights of all the Conv layers from a network trained on the ImageNet dataset[37]. The two FC layers were initialised from random noise.

In contrast to the FC layers, the weights of the Conv layers were not trained during the first 2 epochs (full passes through the training set) and then they were unfrozen. Subsequently, all the layers of the CNN were trained for 20 epochs. Such strategy ensured that the FC layers did not corrupt the pre-trained Conv weights during the first backpropagation passes. The CNN was trained with a learning rate of $1e-3$ (dropped at 15^{th} epoch), batch size of 64, weight decay of $1e-4$ and Adam optimization method[38]. We also placed a dropout layer[39] with the rate of $p = 0.5$ before each FC layer.

During the training of the CNN, we used random noise addition, random rotation ± 5 degrees, random cropping of the original 310×310 pixels image to 300×300 pixels (135×135 mm) and also random gamma correction. These data augmentations were performed randomly on-the-fly, with the aim to train our model to be invariant towards different data acquisition parameters. We used the SOLT package of version 0.1.3[40] in our experiments.

Inference pipeline. At the test phase, we averaged the outputs of all the models trained in cross-validation. Additionally, for each CNN model here, we performed 5-crop test-time augmentation (TTA). Specifically, we cropped 4 images of 300×300 pixels from the corners of the original image, and one same-sized crop from the centre of the image. The predictions for the 5 cropped images were eventually averaged. Subsequently, having the TTA prediction for each cross-validation model, we averaged their results as well. This approach allowed us to reduce the variance of the CNNs and boost the prediction accuracy.

It is worth to mention that during the evaluation of CNN model alone, instead of using the fine-grained division into progression classes, we used the probability of progression $P(prog|x)$ as a sum of $P(y = 1|x)$ and $P(y = 2|x)$. A similar technique was previously utilised in a skin cancer prediction study[41].

Interpreting neural network's decisions. In this study, we focused not only on producing the first state-of-the-art model for knee OA progression prediction, but also developed an approach to examine the network's decision to assess the radiological features detected by the network. Similar to our previous study[7], we modified the GradCAM method[24] to operate with TTA. The output of the GradCAM is an attention map, showing which region of the image positively correlates with the output of the network.

In the previous section, we described a TTA-approach and it should be noted that all the operations including the sum of the progression probabilities are fully differentiable, thus the application of the GradCAM here is fairly straightforward.

Model stacking: fusing heterogeneous data using tree gradient boosting. We fused the predictions of the neural network – KL grade and progression probabilities $P(KL = i|x), i \in \{0, \ldots, 4\}$ and $P(y = i|x)$, $i \in \{0, 1, 2\}$ respectively – with other clinical measures such as patient's age, sex, BMI, previous injury history, symptomatic assessments (WOMAC) and, optionally, a KL grade. Such fusion is challenging, prone to overfitting and requires a robust cross-validation scheme. A stacked generalisation approach, proposed by Wolpert[25] allows to build multiple layers of models and handle these issues.

Following our model inference strategy, we first trained the 5 CNN models corresponding to the 5 cross-validation train-validation splits. Subsequently, this allowed to perform the inference on each validation set in our cross-validation setup and, therefore, obtain CNN predictions for the whole training set. When building the second-level GBM, we utilised the same cross-validation split and used the predictions for each knee joint as input features, along with the other clinical measures.

Statistical analyses. We utilised Precision-Recall (PR) and ROC curves as the main methods to measure the performance of all the methods. PR curve can be quantitatively summarised using the AP metric. The AP metric gives a general understanding on average positive predictive value (PPV) of the method. PPV indicates the probability of the object predicted as positive (progressor in the case of this study) actually being positive. The precision-recall curve has been shown to be more informative than the ROC curve when comparing classifiers on imbalanced datasets[42]. ROC curve can quantitatively be summarised using the AUC. ROC curve demonstrates a trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) of the classifier. AUC represents the quality of ranking random positive examples over the random negative examples[43].

To compute the AUC and AP on the test set, we used stratified bootstrapping with 2,000 iterations. The stratification allowed us to reliably assess the confidence intervals for both AUC and AP. We assessed the statistical significance of the difference between the models using DeLong's test[44].

Data availability

OAI and MOST datasets are publicly available datasets and can be requested at http://most.ucsf.edu/ and https://oai.epi-ucsf.org/. The Dockerfile, source codes, pre-trained models and other relevant data are publicly available at https://github.com/MIPT-Oulu/OAProgression.

Received: 23 May 2019; Accepted: 6 December 2019;
Published online: 27 December 2019

References

1. Arden, N. & Nevitt, M. C. Osteoarthritis: epidemiology. *Best practice & research Clinical rheumatology* **20**, 3–25 (2006).
2. Ferket, B. S. *et al*. Impact of total knee replacement practice: cost effectiveness analysis of data from the osteoarthritis initiative. *bmj* **356**, j1131 (2017).
3. Bedson, J., Jordan, K. & Croft, P. The prevalence and history of knee osteoarthritis in general practice: a case–control study. *Family practice* **22**, 103–108 (2005).
4. Jamshidi, A., Pelletier, J.-P. & Martel-Pelletier, J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews Rheumatology* 1 (2018).
5. van Oudenaarde, K. *et al*. General practitioners referring adults to mr imaging for knee pain: a randomized controlled trial to assess cost-effectiveness. *Radiology* **288**, 170–176 (2018).
6. Kellgren, J. & Lawrence, J. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases* **16**, 494 (1957).
7. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P. & Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific reports* **8**, 1727 (2018).
8. Norman, B., Pedoia, V., Noworolski, A., Link, T. M. & Majumdar, S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *Journal of digital imaging* 1–7 (2018).
9. Antony, J., McGuinness, K., O'Connor, N. E. & Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, 1195–1200 (IEEE, 2016).
10. Kerkhof, H. J. *et al*. Prediction model for knee osteoarthritis incidence, including clinical, genetic and biochemical risk factors. *Annals of the rheumatic diseases* **73**, 2116–2121 (2014).
11. Janvier, T. *et al*. Subchondral tibial bone texture analysis predicts knee osteoarthritis progression: data from the osteoarthritis initiative: tibial bone texture & knee oa progression. *Osteoarthritis and cartilage* **25**, 259–266 (2017).
12. Janvier, T., Jennane, R., Toumi, H. & Lespessailles, E. Subchondral tibial bone texture predicts the incidence of radiographic knee osteoarthritis: data from the osteoarthritis initiative. *Osteoarthritis and cartilage* **25**, 2047–2054 (2017).
13. Kraus, V. B. *et al*. Trabecular morphometry by fractal signature analysis is a novel marker of osteoarthritis progression. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* **60**, 3711–3722 (2009).
14. Yu, D. *et al*. Development and validation of prediction models to estimate risk of primary total hip and knee replacements using data from the uk: two prospective open cohorts using the uk clinical practice research datalink. *Annals of the rheumatic diseases* **78**, 91–99 (2019).
15. Hosnijeh, F. S. *et al*. Development of a prediction model for future risk of radiographic hip osteoarthritis. *Osteoarthritis and cartilage* **26**, 540–546 (2018).
16. Emrani, P. S. *et al*. Joint space narrowing and kellgren–lawrence progression in knee osteoarthritis: an analytic literature synthesis. *Osteoarthritis and Cartilage* **16**, 873–882 (2008).
17. LaValley, M. P., McAlindon, T. E., Chaisson, C. E., Levy, D. & Felson, D. T. The validity of different definitions of radiographic worsening for longitudinal studies of knee osteoarthritis. *Journal of clinical epidemiology* **54**, 30–39 (2001).
18. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, Published online 2014; based on TR arXiv:1404.7828 [cs.NE] (2015).
19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436 (2015).
20. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232 (2001).
21. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*. (Springer series in statistics, New York, 2001).
22. Bellamy, N., Buchanan, W. W., Goldsmith, C. H., Campbell, J. & Stitt, L. W. 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. *The Journal of rheumatology* **15**, 1833–1840 (1988).
23. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
24. Selvaraju, R. R. *et al*. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626 (2017).
25. Wolpert, D. H. Stacked generalization. *Neural networks* **5**, 241–259 (1992).
26. Olah, C. *et al*. The building blocks of interpretability. *Distill* **3**, e10 (2018).
27. Bach, S. *et al*. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **10**, e0130140 (2015).
28. Kinds, M. B. *et al*. Quantitative radiographic features of early knee osteoarthritis: development over 5 years and relationship with symptoms in the check cohort. *The Journal of rheumatology* **40**, 58–65 (2013).
29. Neumann, G. *et al*. Location specific radiographic joint space width for osteoarthritis progression. *Osteoarthritis and cartilage* **17**, 761–765 (2009).
30. Kothari, M. *et al*. Fixed-flexion radiography of the knee provides reproducible joint space width measurements in osteoarthritis. *European radiology* **14**, 1568–1573 (2004).
31. Lindner, C., Bromiley, P. A., Ionita, M. C. & Cootes, T. F. Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence* **37**, 1862–1874 (2015).
32. Pedregosa, F. *et al*. Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).
33. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, vol. 57, 61 (Scipy, 2010).
34. Ke, G. *et al*. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 3146–3154 (2017).
35. Paszke, A. *et al*. Automatic differentiation in pytorch. In *NIPS-W* (2017).
36. Bergstra, J., Yamins, D. & Cox, D. D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, 13–20 (Citeseer, 2013).
37. Deng, J. *et al*. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
38. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
39. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**, 1929–1958 (2014).

40. Tiulpin, A. Solt: Streaming over lightweight transformations, https://github.com/MIPT-Oulu/solt (2019).
41. Esteva, A. *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115 (2017).
42. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, e0118432 (2015).
43. Cortes, C. & Mohri, M. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, 313–320 (2004).
44. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

Acknowledgements

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. 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 strategic funding of the University of Oulu, Infotech Oulu, KAUTE foundation and Sigrid Juselius Foundation for supporting this work. Dr. Claudia Lindner is acknowledged for providing BoneFinder and Egor Panfilov is acknowledged for proof-reading of the manuscript.

Author contributions

A.T. and S.S. originated the idea of the study. A.T., S.S. and S.K. designed the study, A.T. performed the experiments and wrote the manuscript S.K., J.T. and E.R. provided the technical feedback. S.B., E.O. and J.M. provided the clinical feedback. All authors participated in the manuscript writing and editing.

Competing interests

Mr. Aleksei Tiulpin is a co-founder and a shareholder of Ailean Technologies Oy. Other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-56527-3.

Correspondence and requests for materials should be addressed to A.T.