Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks

Kevin A. Thomas, BSE • Łukasz Kidziński, PhD • Eni Halilaj, PhD • Scott L. Fleming, MS • Guhan R. Venkataraman, BS • Edwin H. G. Oei, MD, PhD • Garry E. Gold, MD, MS • Scott L. Delp, PhD

From the Departments of Biomedical Data Science (K.A.T., S.L.F., G.R.V.), Bioengineering (Ł.K., S.L.D.), and Radiology (G.E.G.), Stanford University, Clark Center, 318 Campus Dr, Room S321, Stanford, CA 94305; Department of Radiology, Erasmus University Rotterdam, Rotterdam, the Netherlands (E.H.G.O.); and Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pa (E.H.). Received April 24, 2019; revision requested May 29; revision received October 30; accepted November 6. **Address correspondence to** K.A.T. (e-mail: *kevin.a.thomas@stanford.edu*).

Work supported by NIH Big Data to Knowledge (BD2K) Research Grant U54EB020405

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(2):e190065 • https://doi.org/10.1148/ryai.2020190065 • Content codes: IN MK

Purpose: To develop an automated model for staging knee osteoarthritis severity from radiographs and to compare its performance to that of musculoskeletal radiologists.

Materials and Methods: Radiographs from the Osteoarthritis Initiative staged by a radiologist committee using the Kellgren-Lawrence (KL) system were used. Before using the images as input to a convolutional neural network model, they were standardized and augmented automatically. The model was trained with 32 116 images, tuned with 4074 images, evaluated with a 4090-image test set, and compared to two individual radiologists using a 50-image test subset. Saliency maps were generated to reveal features used by the model to determine KL grades.

Results: With committee scores used as ground truth, the model had an average F1 score of 0.70 and an accuracy of 0.71 for the full test set. For the 50-image subset, the best individual radiologist had an average F1 score of 0.60 and an accuracy of 0.60; the model had an average F1 score of 0.64 and an accuracy of 0.66. Cohen weighted κ between the committee and model was 0.86, comparable to intraexpert repeatability. Saliency maps identified sites of osteophyte formation as influential to predictions.

Conclusion: An end-to-end interpretable model that takes full radiographs as input and predicts KL scores with state-of-the-art accuracy, performs as well as musculoskeletal radiologists, and does not require manual image preprocessing was developed. Saliency maps suggest the model's predictions were based on clinically relevant information.

Supplemental material is available for this article.

Knee osteoarthritis (OA) is a leading cause of disability in older adults, with no effective treatments currently available and dissatisfaction rates of nearly 20% in patients who undergo joint replacement surgery (1). To facilitate the development of treatments, there is a need to make disease staging more efficient. Several methods currently exist for OA staging. Radiographically derived evaluations, particularly joint space narrowing measurements and the Kellgren-Lawrence (KL) scoring system (2) (Fig 1) are some of the most common. Evaluations derived from MRI, such as cartilage volume, morphology, and T2 mapping, have also been shown to provide sensitive measurements of OA worsening. The KL scoring system is particularly valuable in research, where it is often used to define study cohorts and to evaluate how the radiographic manifestation of OA relates to clinical outcomes. For example, previous studies have shown that presurgical KL grade is predictive of surgical success (3). However, interclinician agreement and intraclinician repeatability are not optimal (4,5). More accurate, consistent diagnosis of OA stage would ensure that investigative treatments are evaluated in patients in the severity range intended by the investigators. Furthermore, a time- and cost-efficient approach would accelerate clinical trials, which are often slowed by their reliance on experts to first screen large populations with radiography to identify patients with the appropriate level of OA severity to be included. In clinical practice, having a consistent, automated mechanism for evaluating sequential radiographic examinations of individual patients would enable better tracking of their disease progression.

Given the importance of radiographic staging, automated tools that mitigate human bias and costs are needed. Previous studies have developed automated radiographic classifiers, ranging from machine learning approaches that rely on image information distilled using domain knowledge from computer vision experts (6,7) to deep learning approaches (8,9). These studies have been made possible by the Osteoarthritis Initiative (OAI), in which thousands of knee radiographs have been graded by a radiologist committee. However, methods that rely on image information distilled using domain knowledge from experts have limited accuracy. Deep learning models have performed better, but previous approaches have relied on annotations of the joint space (ie, isolating the small portion of the image containing the joint space). Manual methods for this are time-consuming and potentially noisy, whereas automated

Abbreviations

CI = confidence interval, KL = Kellgren-Lawrence, OA = osteoarthritis, OAI = Osteoarthritis Initiative

Summary

An end-to-end interpretable model that takes full knee radiographs as input and assesses osteoarthritis severity with comparable performance to individual musculoskeletal radiologists and does not require manual image preprocessing was developed.

Key Points

- Our model's Kellgren-Lawrence (KL) scoring agrees with a committee of musculoskeletal radiologists as closely as the best individual musculoskeletal radiologists agree with themselves.
- Our model detects the presence of radiographic osteoarthritis (KL $\geq$ 2) as accurately as musculoskeletal radiologists and is freely available at *https://simtk.org/projects/oastaging/*.
- It is robust to variability in image contrast, joint size, joint location in the frame, and limb side; it automatically detects the relevant regions of the image instead of requiring a user to manually pick out the joint space for the model, and it can assess both right and left limbs without requiring users to first manually mirror all left limbs to look like right limbs.

methods have been found to incorrectly localize the joint on some images.

The aim of this study was to develop a fully automated model for staging knee OA severity and to compare its performance to fellowship-trained musculoskeletal radiologists' performance. Because the model is end-to-end and takes as input the same full-sized radiograph as that viewed by the radiologist, comparisons with radiologists have immediate relevance. Here we take a deep learning approach, training a convolutional neural network model. By utilizing data augmentation, we eliminate the need to manually annotate images during both the training and application of the model. We have made our model publicly available at *https://simtk.org/projects/oastaging/*.

Materials and Methods

We used one set of images to train a neural network model to identify radiographic features that were indicative of OA severity to make accurate KL predictions and then used a different, held-out set of images to assess how well the model could assess KL scores for previously unseen images. The held-out set of images used to evaluate the model (ie, the test set) was not used to train the model. To maximize the model's ability to generalize to new images, each training image was replicated into several altered versions to increase the effective diversity within our training set. When the model made KL predictions on the test set, we assessed which regions of each image were most influential for the prediction to determine whether the model was "paying attention" to the same radiographic features as radiologists.

Dataset

The OAI is a longitudinal observational study of knee OA available for public access (10). More than 4000 men and women were assessed annually for nearly a decade. Bilateral fixed flexion plain film radiographs were taken, and each joint of each radiograph was staged using the KL system by two trained musculoskeletal radiologists. Disagreements were resolved by a third radiologist, and their consensus score was reported. We obtained institutional review board approval from our institution to carry out this investigation with the OAI data.

We used six longitudinal bilateral radiographic images of 4508 patients, yielding 40 280 single-limb images (ie, 20 140 total bilateral images) despite missing data for some patients at some time points. Reasons for missing data include patients withdrawing from the study or dying. Patients were randomly split into training (3606 patients, 32 116 single-limb images), validation (450 patients, 4074 single-limb images), and test (452 patients, 4090 single-limb images) sets. Women (58% of patients) had an average age of 60.9 years (age range, 45–79 years). Men (42% of patients) had an average age of 61.3 years (age range, 45–79 years). All images from a specific patient were included only in one set.

Individual Radiologist Evaluations

To measure agreement between the OAI committee, model, and individual radiologists, KL scores were collected from two musculoskeletal subspecialty radiologists (E.H.G.O. and G.E.G.) for 50 of the test set images—10 randomly chosen radiographs from each KL grade. The radiographs were provided to the radiologists in the OAI's original, high-resolution format. Both radiologists were blinded to OAI committee scores and the scores of the other radiologist. Both radiologists had several years of experience in applying the KL scoring system and had examined numerous OAI radiographs and their corresponding OAI committee KL scores through past research. G.E.G. previously served on the Imaging Advisory Board for the OAI. They were therefore well calibrated to the OAI committee's KL scoring tendencies when conducting their own evaluations of this test set.

Model Architecture

A 169-layer convolutional neural network with a dense convolutional network architecture (Fig 2) was used to predict the KL score for each image (11). This architecture has shown success in other orthopedic radiograph classification tasks, including tasks comparable to KL scoring in which only a small portion of each overall image may be relevant for determining class assignments (12). The final layer was modified to have five outputs, one for each KL class. The weights of the network were initialized with weights from a model pretrained on ImageNet, a large annotated database used to train computer vision models. A softmax nonlinearity function (13) was then applied over the five outputs to convert them into the probabilities that a given image represents each of the five KL scores. The model was trained to produce predictions that minimize the cross entropy between the OAI committee's scores and its own predicted scores (Appendix E1 [supplement]). As part of our hyperparameter search, we also evaluated the Inception v3 model architecture (14). However, this provided consistently lower performance on the validation set and was therefore not

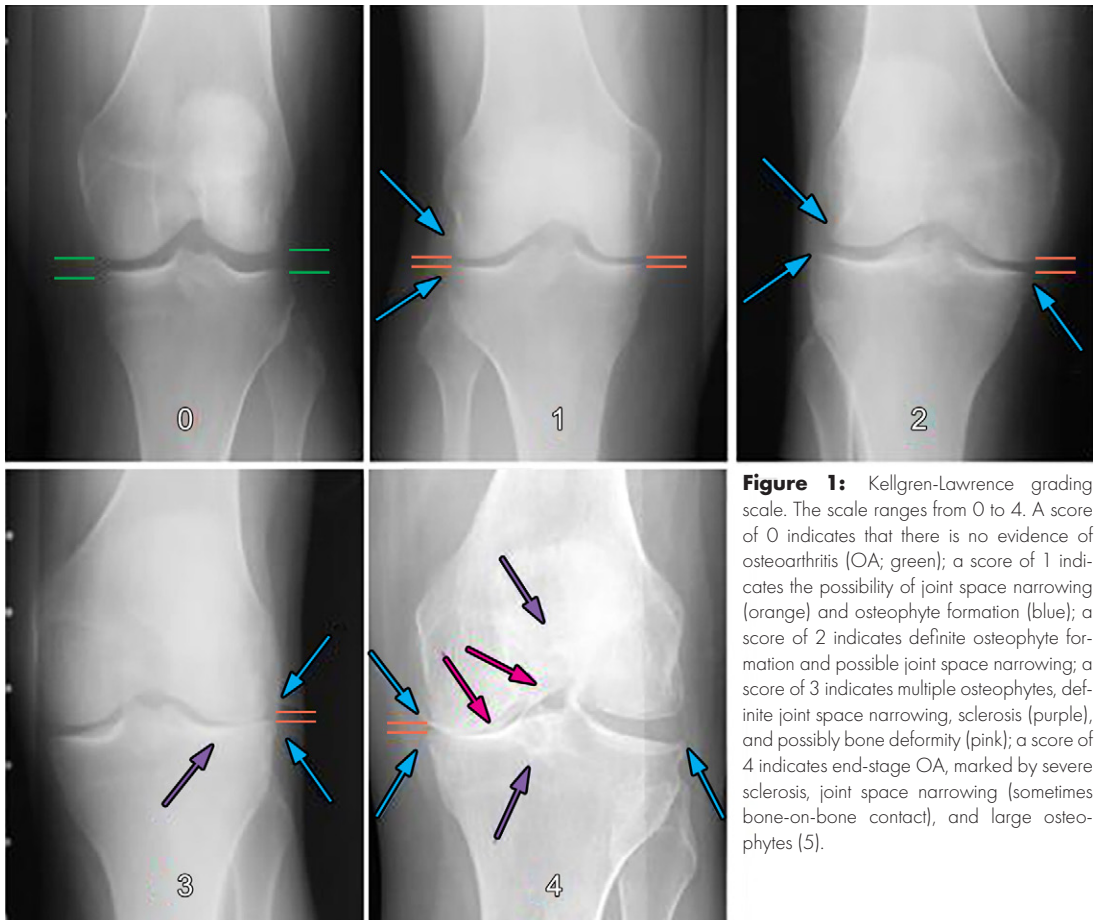

Figure 1: Kellgren-Lawrence grading scale. The scale ranges from 0 to 4. A score of 0 indicates that there is no evidence of osteoarthritis (OA; green); a score of 1 indicates the possibility of joint space narrowing (orange) and osteophyte formation (blue); a score of 2 indicates definite osteophyte formation and possible joint space narrowing; a score of 3 indicates multiple osteophytes, definite joint space narrowing, sclerosis (purple), and possibly bone deformity (pink); a score of 4 indicates end-stage OA, marked by severe sclerosis, joint space narrowing (sometimes bone-on-bone contact), and large osteophytes (5).

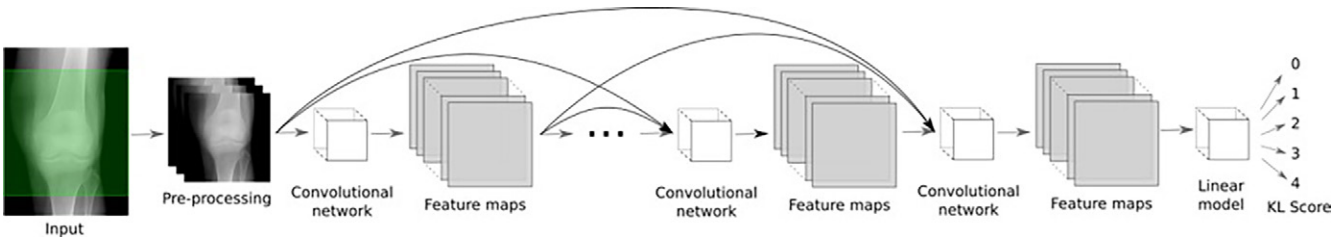

Figure 2: DenseNet architecture. Deep convolutional neural networks are composed of a sequence of sets of convolutional filters with parameters trained from the data. In DenseNet, every pair of layers is connected so that low-level features from the first layers (such as edges and primitive shapes) can be used directly in the fully connected layer (ie, linear model) of the network. KL = Kellgren-Lawrence.

included in final analyses. All modeling and analyses were done using the Python programming language, version 3.5. Original model architectures were obtained from the TorchVision package available in Python before being modified. Models were trained and evaluated using one NVIDIA K80 GPU (Santa Clara, Calif).

Data Preprocessing and Augmentation

Each image was split down the middle to produce right and left knee images. All images were resampled to 299 × 299 pixels because our neural network model requires all input images to have a consistent size, and 299 × 299 is the higher default resolution of the two model architectures we initially considered. Because the original images varied in resolution and zoom, pixel size differed between images. The default preprocessing for all single-joint images across training, validation, and test sets involved the procedure detailed on the left in Table 1. During training, the default preprocessing was used with 20% probability for each image in each epoch (ie, each round of training in which a model sees every training image once) (Fig 3). To augment the dataset, training images were augmented using the procedure detailed on the right in Table 1, with 80% probability in each epoch (Fig 3). Augmentation was motivated by the observation that OAI images vary substantially in image contrast, position of the joint space within the image frame, relative zoom, and side (ie, right or left limb). By replicating each training set image into several altered versions (eg, mirroring an image and changing

Table 1: Image Preprocessing Procedure

Step	Default Preprocessing Procedure	Augmentation Preprocessing Procedure
1	Remove an equal number of rows from the top and bottom of the image to obtain a square image	Randomly imbalance the number of rows cropped from the top versus bottom of the image during the original cropping such that the final number of removed rows remains the same as in the original procedure. The ratio of rows removed from the top versus bottom was a uniformly distributed random variable ranging from 1:3 to 3:1. A new ratio was selected from this distribution for each augmented training image in each epoch. Random additional cropping to obtain a square subset of the image obtained from the above step. The percentage of rows and columns retained was a uniformly distributed random variable ranging from 70% to 95% of the number of rows and columns obtained from the previous set.
2	Resample with cubic interpolation to obtain a 299 × 299 pixel image	Resample with cubic interpolation to obtain a 299 × 299 pixel image
3	Center pixel values using the individual image's mean value	Center pixel values using the individual image's mean value
4	Scale pixel values using the individual image's standard deviation	Scale pixel values using the individual image's standard deviation Add Gaussian noise with mean 0 and a randomly selected standard deviation between 0 and 0.1
5	Rescale such that the largest pixel value was 1 and the smallest pixel value was 0	Rescale image such that the largest pixel value was 1 and the smallest pixel value was 0 Multiply pixel values by a random number between 0.6 and 2.0, then subtract 1 from each pixel Mirror along the vertical axis with a probability of 50%
6	Replicate image into three channels	Replicate into three channels
7	Center using ImageNet RGB channel means	Center using ImageNet RGB channel means
8	Scale using ImageNet channel standard deviations	Scale using ImageNet channel standard deviations

Note.—The default preprocessing procedure followed a deterministic approach, whereas the augmented preprocessing procedure took a probabilistic approach so that no two augmented versions of a given image were the same. RGB = red, green, blue.

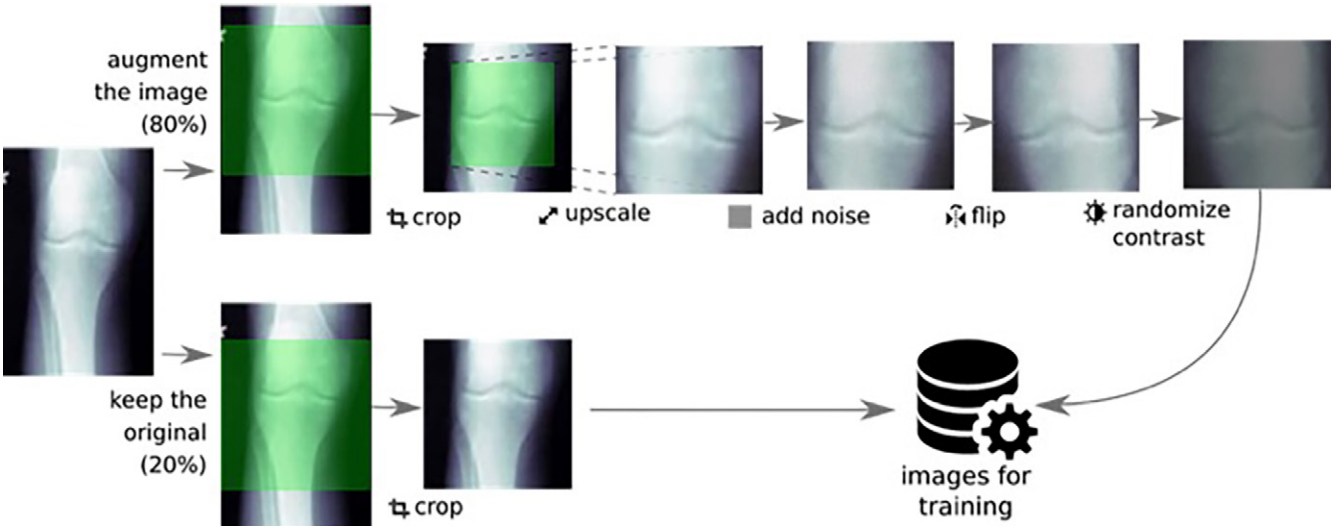

Figure 3: Data augmentation illustration. In each epoch, the original image is either preprocessed with the standard procedure (20% probability) or it undergoes the data augmentation procedure (80% probability). In the data augmentation, we crop, zoom in, upscale, add noise, flip horizontally, and adjust contrast in a stochastic manner so that the generated images follow the distribution of images in the original dataset.

its contrast to convert a right knee with high contrast into a left knee with lower contrast), we better prepared the model to make predictions for new images with different combinations of these parameters than were found in the original, nonaugmented training set.

A new random augmentation procedure was performed with each epoch and each image. One original image was converted into a different augmented image with each epoch, and no two images were augmented with the same augmentation parameters (Fig 4). To examine the value of our augmentation procedure, an

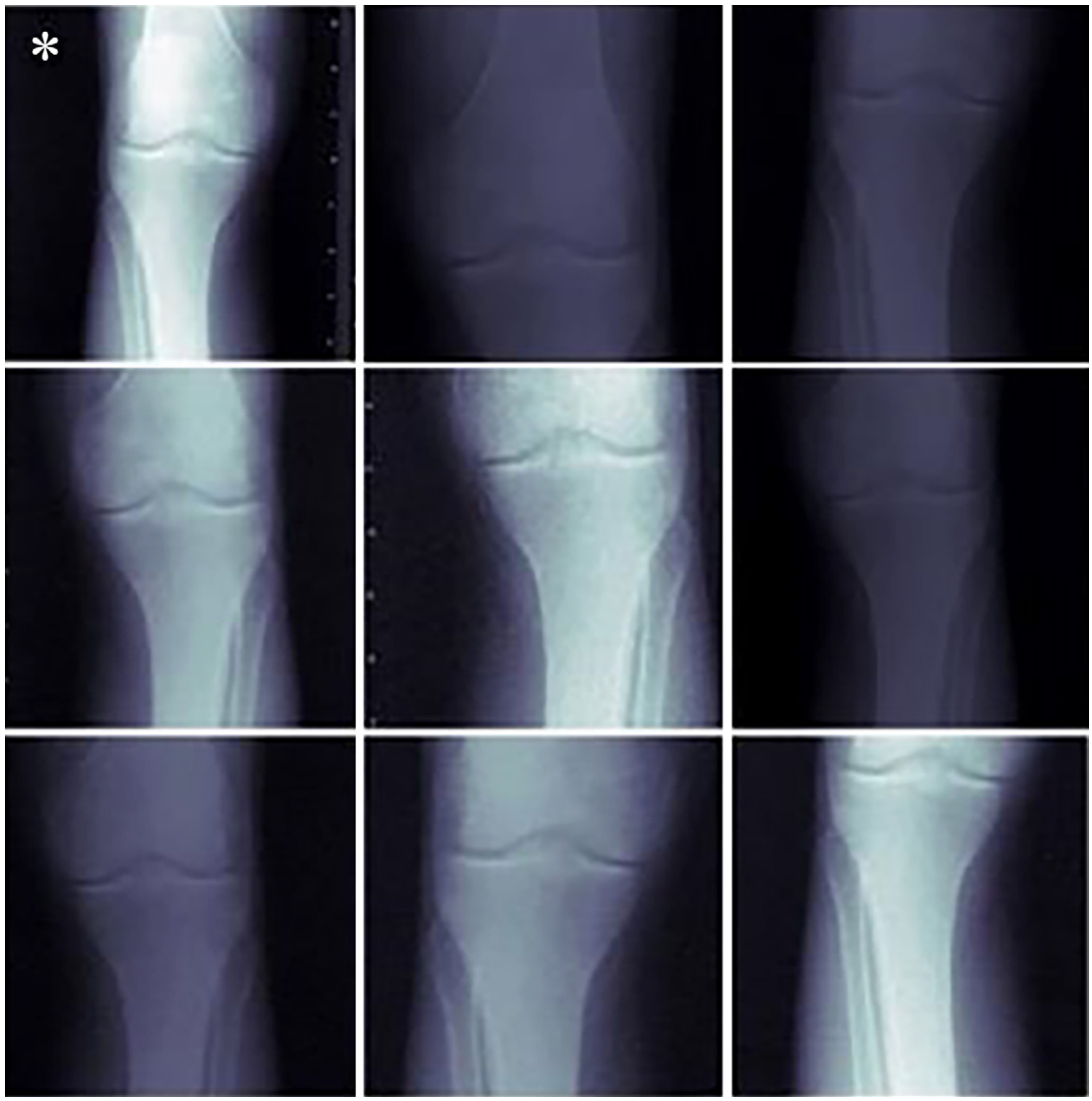

Figure 4: Example of data augmentation. Original image is shown in top left corner (marked with *). Contrast, zoom, and position of joint were randomly varied, and the image was randomly mirrored. The augmented data were used to improve the model's ability to generalize to new images.

additional model was trained without the use of augmentation and assessed with the same test set as the model that was trained with augmentation.

Performance Evaluation and Comparison with Radiologist

We used three metrics for evaluation: accuracy, F1 score, and Cohen weighted κ. Although intuitive, accuracy is misleading for this dataset because there is an unequal number of images representing each KL score. We used the F1 score and κ (15) because both are robust to imbalance in the number of images with each KL grade and both have been used in other studies on KL scoring, enabling comparison. The F1 score is a single metric that combines precision and recall and has a range of 0–1, where 1 denotes perfect agreement. For multiclass classification, we compute F1 scores for each class separately and average the results. F1 score confidence intervals (CIs) were obtained by bootstrapping 10 000 simulated test set confusion matrices from either our model's or other publications' test set confusion matrix, calculating performance metrics for each simulated confusion matrix, and then calculating the 95% CI of the performance metrics. κ evaluates agreement between two labeling methods without specifying one as ground truth. κ values $\leq$ 0 indicate no agreement, 0.5 denotes moderate agreement, and 1.0 denotes perfect agreement. κ accounts for the magnitude of disagreement between two evaluators and the likelihood that they agree by random chance, whereas F1 only considers whether an evaluator is correct.

Model Interpretation

Saliency maps were used to obtain a qualitative understanding of how the trained model arrived at predictions (Fig 5). They were produced by calculating the contribution of each pixel to the probability that the model assigns to the true KL class. This was done by performing backpropagation through the trained model from the single probability output value assigned to the true KL class back to the pixels of the image that produced that probability.

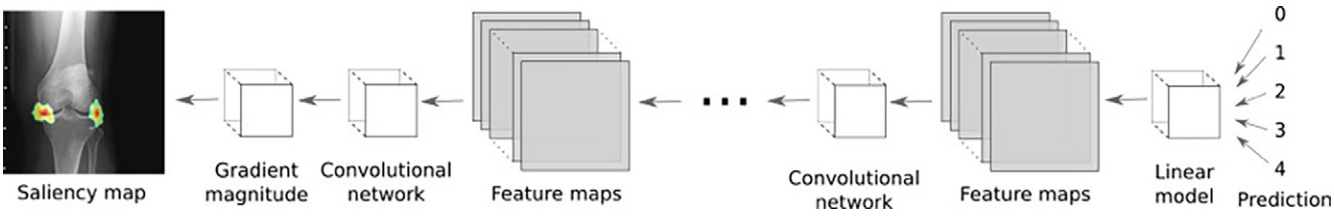

Figure 5: Saliency map algorithm. Contribution of each pixel is derived from the backpropagation process. Variability in the output layer is passed through the network and exposes pixels that contributed most to this variability. Changes in the input image in these pixels would affect the predicted score the most. We interpret these pixels as the most predictive ones. Here, we represent this intensity mapping with an image transparent for very low values, green for low values, and red for large values. This intensity matrix is referred to as a saliency map.

Table 2: Radiologist's and Model's Agreement with OAI Committee

	Best Radiologist (E.H.G.O.) 50-Image Subset			Model 50-Image Subset			Model Full Test Set			Antony et al (8)		
Parameter	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
KL score												
0	0.71	0.50	0.59	0.53	1.00	0.69	0.73	0.87	0.79	0.57	0.92	0.71
1	0.44	0.40	0.42	0.50	0.20	0.29	0.38	0.27	0.31	0.32	0.14	0.20
2	0.50	0.70	0.58	0.60	0.60	0.60	0.71	0.67	0.69	0.71	0.46	0.56
3	0.60	0.60	0.60	0.78	0.70	0.74	0.82	0.81	0.81	0.78	0.73	0.76
4	0.80	0.80	0.80	1.00	0.80	0.89	0.87	0.86	0.87	0.89	0.73	0.80
Mean	0.61 (0.47, 0.75)	0.60 (0.46, 0.73)	0.60 (0.45, 0.72)	0.68 (0.53, 0.83)	0.64 (0.55, 0.77)	0.64 (0.50, 0.76)	0.70 (0.69, 0.72)	0.69 (0.68, 0.71)	0.70 (0.68, 0.71)	0.61*	0.62*	0.59*
Accuracy	0.6 (30/50)			0.66 (33/50)			0.71 (2890/4090)			0.60		

Note.—Precision (Prec), recall, and F1 score for each Kellgren-Lawrence (KL) score and their mean, with 95% confidence interval in parentheses, across all KL scores. Accuracy calculated using all 50 images in test subset or all 4090 images in full test set. Data in parentheses are raw data. OAI = Osteoarthritis Initiative.
* Mean precision, recall, and F1 reported by Antony et al (8) were weighted according to the frequency of each KL score in their sample, whereas our mean metrics were simple averages.

Results

With the committee scores used as ground truth, the model's predictions had a simple average F1 score of 0.70 and an accuracy of 0.71. The model converged after six epochs of training, requiring 18 hours on our system. When we trained a model using only the default preprocessing procedure and did not use augmentation, it achieved a simple average F1 score of 0.66 and an accuracy of 0.68. For comparison, the best-performing model in previous literature reported a class-weighted average F1 score of 0.59 and an accuracy of 0.60 (8). Calculating a simple average F1 score from their single-class data yields 0.61.

For the 50-image test subset that was evaluated by E.H.G.O. and G.E.G., the best average F1 score and overall accuracy between them came from E.H.G.O. and were both 0.60. The model had an average F1 score of 0.64 and an overall accuracy of 0.66 for this test subset. The model's F1 scores for individual KL scores exceeded those of E.H.G.O. for KL of 0, 2, 3, and 4, whereas E.H.G.O. had higher F1 scores for KL of 1 (Table 2). Because the subset contains an equal number of images from each KL class, these results can be directly compared with the weighted F1 scores reported in Antony et al (8).

Norman et al merged the KL of 0 and KL of 1 classes based on the fact that both represent an absence of OA (16). When we merged the KL of 0 and KL of 1 predictions for our model, it had a mean F1 score of 0.80 (95% CI: 0.782, 0.814). For comparison, calculating the mean F1 score from their confusion matrix yielded a mean F1 score of 0.768 (95% CI: 0.753, 0.782). Their primary performance metrics were sensitivity and specificity. Averaging across the four KL classes, our model's mean sensitivity was 0.799 (95% CI: 0.783, 0.816) and its mean specificity was 0.917 (95% CI: 0.911, 0.922). These are comparable with their model's mean sensitivity of 0.772 (95% CI: 0.757, 0.787) and mean specificity of 0.915 (95% CI: 0.911, 0.920). However, our F1 score is significantly higher. Our model's confusion matrix is presented in Table 3. See also Tables E1–E6 (supplement).

The KL of 2 score has special importance because it is often used as the threshold for determining OA incidence when using the KL system for cohort selection. To assess the ability of the model to determine incidence of OA, we combined the 0 and 1 KL scores into one class and combined the 2, 3, and 4 KL scores into another class. For the 50-image test subset that was evaluated by E.H.G.O. and G.E.G., the best average F1 score and accuracy for detecting OA incidence came from E.H.G.O. and were 0.875

Table 3: Confusion Matrix for Full Test Set

		Model's Predictions				
		0	1	2	3	4
OAI scores	0	1247	127	65	2	0
	1	324	177	146	9	0
	2	136	156	744	78	0
	3	0	8	91	534	27
	4	0	0	0	31	188

Note.—The entry in row r and column c denotes the number of images that the Osteoarthritis Initiative committee labeled as Kellgren-Lawrence (KL) = r and our model labeled as KL = c. For example, there are 1247 images that the committee labeled as KL = 0 and the model also labeled as KL = 0. OAI = Osteoarthritis Initiative.

Table 4: Radiologist's and Model's Agreement with OAI Committee for Determining OA Incidence

Parameter	Best Radiologist (E.H.G.O.) 50-Image Subset	Model 50-Image Subset	Model Full Test Set
F1 score	0.875	0.912	0.866
Precision	0.823	0.963	0.884
Recall	0.933	0.867	0.849
Accuracy*	0.840 (42/50)	0.90 (45/50)	0.872 (3568/4090)

Note.—F1 score, precision, recall, and accuracy for detecting incidence of osteoarthritis (OA) (ie, classifying if Kellgren-Lawrence [KL] score ≥ 2) for the best individual radiologist and for the model. The Osteoarthritis Initiative (OAI) committees' scores were used as ground truth.
* Data in parentheses are raw data.

and 0.840, respectively. Our model had an average F1 score of 0.912 and an accuracy of 0.90 for this test subset, exceeding both individual radiologists. For the full test set, our model had an F1 score of 0.866 and an accuracy of 0.872 (Table 4).

The model outputs five probabilities for each image, corresponding to the probability that a given image represents each of the five KL scores. We identified the test set images that E.H.G.O. and G.E.G. incorrectly classified and examined the probability that the model assigned the correct KL score for these images. The model assigned a smaller average probability to the correct KL score for these images (0.46 and 0.56) than the average probability it assigned for images that the radiologists correctly classified (0.71 and 0.64). The individual radiologists' accuracy when they agreed with the model was higher (0.81 and 0.61) than their accuracy when they disagreed with the model (0.34 and 0.41).

When assessing agreement with the OAI committee, the best κ across E.H.G.O. and G.E.G. was 0.86 with the 50-image test subset, 0.90 for the model with the 50-image test subset, 0.86 for the model with the full test set, and 0.83 for Tiulpin et al's model, which has the highest κ in previously published literature (9) (Table 5). When assessing interrater agreement, the κ was 0.81 between E.H.G.O. and the model, 0.89 between G.E.G. and the model, 0.79 between E.H.G.O. and G.E.G., and 0.65 for the two most agreeing raters in Riddle et al (4) (Table 5). When assessing intrarater agreement, the model is guaranteed to have a κ of 1.0. Intrarater agreement was not assessed for E.H.G.O. or G.E.G., but previous work has reported a maximum value of 0.85 for OAI radiographs (4) (Table 5).

Model Interpretation

Saliency maps (Fig 6) showed that the regions of the image containing the medial and lateral joint margins were frequently observed to provide the highest contributions to the model's predictions. The intercondylar tubercles (ie, tibial spines) were also observed to provide salient features to the model. No systematic differences in saliency maps were observed across the KL score spectrum, nor were systematic differences observed between correctly and incorrectly classified images (Figs E1, E2 [supplement]).

Discussion

Our model agrees with the OAI committee with higher κ than the highest reported radiologist intrarater κ in literature and has the additional advantage of a guaranteed intrarater κ of 1.0. These findings suggest that our algorithm approaches the upper bound of possible performance of an experienced radiologist. Directly comparing predictions of the algorithm with annotations of E.H.G.O. and G.E.G. yielded κ of 0.81 and 0.89, which are comparable to the 0.79 κ observed between E.H.G.O. and G.E.G. as well as the maximum intrarater κ of 0.85 reported in Riddle et al (4). This indicates that our model's annotation pattern might be indistinguishable from a human radiologist's annotations. The model could empower individual radiologists to achieve committee-quality evaluation by providing a second assessment, thereby reducing the noise in KL scores.

Prior work on this task has reduced the variability in joint size and joint location across radiographs by cropping images to only include the joint space before using them as input to a model (8,9,16). This step has been deemed important because features relevant to OA are mainly found within the joint space. Human annotation involves manually drawing a box that contains the joint space for each image. Automated methods have been attempted with varying success. One study reported perfect accuracy using a template-based method (10), but another reported low accuracy when implementing it (8). The latter proposed an alternative method using edge-detection features and

reported an average joint detection F1 score of 0.94. Norman et al reported joint mislocalization in 1.7% of OAI test set images using a deep learning model (11). Tiulpin et al (9) reported joint mislocalization in 1.5% of OAI test set images using the algorithm in Tiulpin et al (12) and relied on manual annotation for these cases. The need to annotate images introduces the potential for added noise and error, and it requires additional time. Development of a region proposal model (13,15) that combines automatic joint space identification with KL classification in a single model has been evaluated on portions of the OAI dataset (16), but this approach requires the initial manual annotation of a training set of images. It also potentially makes the model less robust to new datasets. Furthermore, by using shared features to both localize the joint space and classify OA severity, it becomes more difficult to identify the image regions that were used to classify OA severity, thereby reducing model interpretability.

We created a model that was more robust to variability instead of reducing the variability of its input via cropping. We designed a data augmentation process such that the distribution of features of generated images matched the distribution in the original dataset. This probabilistic approach, in which samples are generated ad hoc following the predefined distribution, increased the diversity present within the training set and led the model to learn from several different, modified versions of each image over the course of training. Augmentation also enables the model to work well on both right and left knees. Other works split each bilateral image in half and convert all left legs into right legs via mirror imaging with the justification that it alleviates the need to learn to detect limb side (9,16). However, if given a new image of a single knee, this approach requires manual input to first determine whether it is a left knee that needs to be mirrored or a right knee that should not be mirrored. Augmentation was also observed to improve performance. Training a model with identical DenseNet architecture and hyperparameter search, but using only the default image preprocessing procedure, revealed that our augmentation enhanced the model's performance beyond that of the default preprocessing, increasing the F1 score by 0.04, a 6.1% improvement. Other works have relied on a multimodel approach, instead of augmentation, to obtain an accurate model (9,16). They use one model to crop the joint space and then use an ensemble of other models to classify each image. The use of several models increases the computational time and resources required relative to ours, which uses a single model for the entire analysis.

We provide publicly available software enabling practitioners to analyze their radiographs and receive an automated KL score within 30 seconds with a single CPU and within 2 seconds on a GPU. The data used for training are freely available through the OAI website (10). Our data augmentation algorithm can be extended to new modes of variability, such as rotations or distortions, and the code can be adapted to other radiology tasks of similar structure.

Incorporating the model into research or clinical workflows would be unlikely to add substantial time or labor to the current radiograph collection pipeline. Although clinical decisions regarding joint replacement surgery are usually based on pain and not KL score, research suggesting that presurgical

Table 5: Cohen Weighted κ Comparisons between Raters

Parameter	OAI versus Best Radiologist (E.H.G.O.), 50-Image Subset	OAI versus Model, 50-Image Subset	OAI versus Model, Full Test Set	OAI versus Best Radiologist in Riddle et al (4)	OAI vs Best Past Model, Tiulpin et al (9)	Radiologist versus Model, 50-Image Subset	Best Observed Interrater κ from Riddle et al (4)	Model's Intrarater κ	Best Observed Intrarater κ from Riddle et al (4)
κ	0.86 (0.77, 0.94)	0.90 (0.84, 0.95)	0.86 (0.86, 0.86)	0.80 (0.70, 0.91)	0.83 (0.83, 0.83)	0.89 (0.84, 0.94)	0.65 (0.38, 0.92)	1.0 (1.0, 1.0)	0.85

Note.—The model agrees with human experts as much as they agree amongst themselves. This suggests that the model is approaching the ceiling for predictive performance given the imperfect repeatability observed among human experts. Data in parentheses are confidence intervals. OAI = Osteoarthritis Initiative.

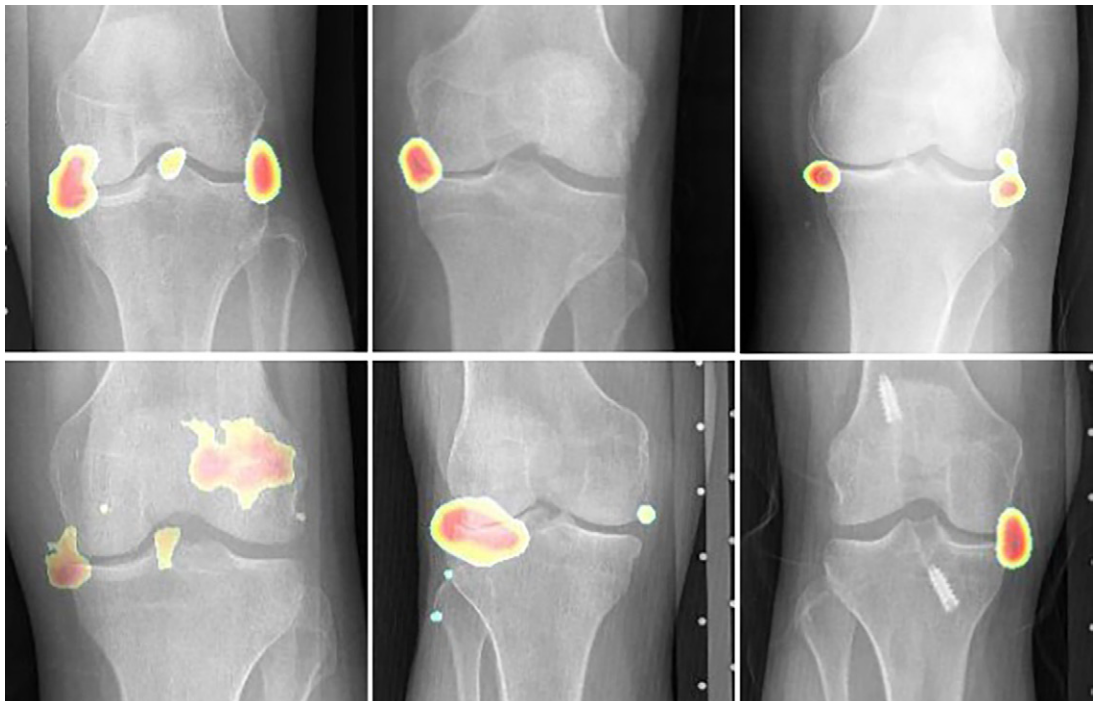

Figure 6: Example saliency maps. Saliency maps were produced and examined for many images in the test set. Medial and lateral joint margins and intercondylar tubercles (ie, tibial spines) were frequently highlighted as important sources of predictive information. These are primary sites of osteophyte formation in osteoarthritis and are indicators used by radiologists in the determination of Kellgren-Lawrence scores. This supports the notion that the model learned clinically relevant features on which to base its predictions.

KL grade is predictive of surgical success (3) supports the idea of using an automated tool like ours to make better-informed decisions. Images can be automatically preprocessed using the default preprocessing described on the left in Table 1 and then classified by the model quickly on a standard computer without GPUs. This could begin immediately after an image is generated without intermediate human involvement, and the prediction could then be made available to the scientist or physician reviewing the image. The software that we have made publicly available performs both the preprocessing and classification, making deployment immediately feasible. However, additional validation studies to establish the accuracy of our model on images outside of the OAI are necessary before the model's KL labels can be used to inform decisions in the clinic and in investigations.

Limitations of the model must be noted. First, we compared its performance to that of radiologists using only 50 images. Although based on a relatively small sample, this comparison provided promising results that have not been previously reported, to our knowledge. Second, the images in the OAI dataset were collected using a standardized protocol. It remains to be seen how the model generalizes to clinical radiographs that may not position the knee in the same way. Joint space narrowing, another measure of OA severity derived from radiographs, is heavily influenced by joint positioning. KL grading may be less dependent on this factor, but this is unknown. Our data augmentation procedure enhances the model's ability to make accurate predictions on new, diverse images.

We have developed an end-to-end model that takes a full knee radiograph as input and predicts the KL score with performance that exceeds other models trained on the OAI dataset and interradiologist agreement. The model we present here is fully automated and provides insight into its decision-making process. We made the model available at *https://simtk.org/projects/oastaging/* as a docker container (17) that can be run with or without GPU acceleration on Windows, Mac OS, and Linux.

Author contributions: Guarantors of integrity of entire study, K.A.T., G.R.V., S.L.D.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, K.A.T., Ł.K., E.H., S.L.F., G.R.V.; clinical studies, E.H., E.H.G.O., G.E.G.; experimental studies, K.A.T., Ł.K., E.H., S.L.F., G.R.V.; statistical analysis, K.A.T., Ł.K., E.H., S.L.F., G.R.V., G.E.G.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: **K.A.T.** disclosed no relevant relationships. **Ł.K.** disclosed no relevant relationships. **E.H.** disclosed no relevant relationships. **S.L.F.** disclosed no relevant relationships. **G.R.V.** disclosed no relevant relationships. **E.H.G.O.** disclosed no relevant relationships. **G.E.G.** Activities related to the present article: institution receives grant from GE Healthcare; author paid consulting fee from Canon Medical. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **S.L.D.** Activities related to the present article: institution received NIH grant. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

References

1. Bourne RB, Chesworth BM, Davis AM, Mahomed NN, Charron KDJ. Patient satisfaction after total knee arthroplasty: who is satisfied and who is not? Clin Orthop Relat Res 2010;468(1):57–63.
2. Kellgren JH, Lawrence JS. Radiological assessment of rheumatoid arthritis. Ann Rheum Dis 1957;16(4):485–493.
3. Mazzuca SA, Brandt KD, Schauwecker DS, et al. Severity of joint pain and Kellgren-Lawrence grade at baseline are better predictors of joint space nar-

rowing than bone scintigraphy in obese women with knee osteoarthritis. J Rheumatol 2005;32(8):1540–1546.
4. Riddle DL, Jiranek WA, Hull JR. Validity and reliability of radiographic knee osteoarthritis measures by arthroplasty surgeons. Orthopedics 2013;36(1):e25–e32.
5. Kessler S, Guenther KP, Puhl W. Scoring prevalence and severity in gonarthritis: the suitability of the Kellgren & Lawrence scale. Clin Rheumatol 1998;17(3):205–209.
6. Shamir L, Ling SM, Scott W, Hochberg M, Ferrucci L, Goldberg IG. Early detection of radiographic knee osteoarthritis using computer-aided analysis. Osteoarthritis Cartilage 2009;17(10):1307–1312.
7. Shamir L, Rahimi S, Orlov N, Ferrucci L, Goldberg IG. Progression analysis and stage discovery in continuous physiological processes using image computing. EURASIP J Bioinform Syst Biol 2010;2010:107036.
8. Antony J, McGuinness K, Connor NEO, Moran K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. ArXiv 1609.02469 [preprint] https://arxiv.org/abs/1609.02469. Posted September 8, 2016. Accessed May 1, 2018.
9. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. Sci Rep 2018;8(1):1727.
10. The Osteoarthritis Initiative. https://oai.epi-ucsf.org/datarelease/. Published 2013. Accessed January 13, 2018.
11. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, July 21–26, 2017. Piscataway, N.J.: IEEE, 2017; 2261–2269.
12. Rajpurkar P, Irvin J, Bagul A, et al. Mura dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. ArXiv 1712.06957 [preprint] https://arxiv.org/abs/1712.06957. Posted December 11, 2017. Accessed November 1, 2019.
13. Bishop CM. Pattern recognition and machine learning. New York, N.Y.: Springer-Verlag, 2006.
14. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 27–30, 2016. Piscataway, N.J.: IEEE, 2016; 2818–2826.
15. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213–220.
16. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J Digit Imaging 2019;32(3):471–477.
17. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J, 2014.