ACR Open Rheumatology
Vol. 6, No. 6, June 2024, pp 388–395
DOI 10.1002/acr2.11665
© 2024 The Authors. *ACR Open Rheumatology* published by Wiley Periodicals LLC on behalf of American College of Rheumatology.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
American College of Rheumatology
Empowering Rheumatology Professionals

Detection and Grading of Radiographic Hand Osteoarthritis Using an Automated Machine Learning Platform

Leo Caratsch,[1,2,3] Christian Lechtenboehmer,[1,2,3,4] Matteo Caorsi,[3] Karine Oung,[1] Fabio Zanchi,[1] Yasser Aleman,[1] Ulrich A. Walker,[4] Patrick Omoumi,[1] and Thomas Hügle[1]

Objective. Automated machine learning (autoML) platforms allow health care professionals to play an active role in the development of machine learning (ML) algorithms according to scientific or clinical needs. The aim of this study was to develop and evaluate such a model for automated detection and grading of distal hand osteoarthritis (OA).

Methods. A total of 13,690 hand radiographs from 2,863 patients within the Swiss Cohort of Quality Management (SCQM) and an external control data set of 346 non-SCQM patients were collected and scored for distal interphalangeal OA (DIP-OA) using the modified Kellgren/Lawrence (K/L) score. Giotto (Learn to Forecast [L2F]) was used as an autoML platform for training two convolutional neural networks for DIP joint extraction and subsequent classification according to the K/L scores. A total of 48,892 DIP joints were extracted and then used to train the classification model. Heatmaps were generated independently of the platform. User experience of a web application as a provisional user interface was investigated by rheumatologists and radiologists.

Results. The sensitivity and specificity of this model for detecting DIP-OA were 79% and 86%, respectively. The accuracy for grading the correct K/L score was 75%, with a κ score of 0.76. The accuracy per DIP-OA class differed, with 86% for no OA (defined as K/L scores 0 and 1), 71% for a K/L score of 2, 46% for a K/L score of 3, and 67% for a K/L score of 4. Similar values were obtained in an independent external test set. Qualitative and quantitative user experience testing of the web application revealed a moderate to high demand for automated DIP-OA scoring among rheumatologists. Conversely, radiologists expressed a low demand, except for the use of heatmaps.

Conclusion. AutoML platforms are an opportunity to develop clinical end-to-end ML algorithms. Here, automated radiographic DIP-OA detection is both feasible and usable, whereas grading among individual K/L scores (eg, for clinical trials) remains challenging.

INTRODUCTION

Digital transformation, enabled by advances in computer performance, data storage, and interoperable user interfaces, has made significant inroads into health care. Deep learning and convolutional neural networks (CNNs) allow the exploitation of medical data on a higher level and provide new forms of clinical decision support.[1] Some of the first artificial intelligence (AI) applications to receive US Food and Drug Administration approval in rheumatology were algorithms for image recognition of radiographs, such as in knee osteoarthritis (OA), which are on the market and available for routine clinical practice.[2] Applications for the automated radiographic detection of inflammatory lesions, such as in rheumatoid arthritis (RA), are also in development.[3]

No models currently address hand OA at the joint level, such as distal interphalangeal (DIP) joints,[4] a common condition without an effective treatment.[5] Üreten et al developed a classification model that can identify the presence of OA or RA on a whole-hand radiograph.[6] However this model does not identify the presence of OA at individual joint level and cannot tell which joint is affected.

The radiographic evolution of DIP-OA is complex because of its multiphasic and potentially erosive course.[7] Instant diagnostic support for interpreting radiographic images of hand OA could

[1] Leo Caratsch, MD, Christian Lechtenboehmer, MD, Karine Oung, MD, Fabio Zanchi, MD, Yasser Aleman, PhD, Patrick Omoumi, MD, PhD, Thomas Hügle, MD, PhD, MA: Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; [2] Leo Caratsch, MD, Christian Lechtenboehmer, MD: City Hospital Waid, Zurich, Switzerland; [3] Leo Caratsch, MD, Christian Lechtenboehmer, MD, Matteo Caorsi, PhD: L2F (Learn to Forecast), Lausanne, Switzerland; [4] Christian Lechtenboehmer, MD, Ulrich A. Walker, MD, PhD: University Hospital of Basel, Basel, Switzerland.

Additional supplementary information cited in this article can be found online in the Supporting Information section (http://onlinelibrary.wiley.com/doi/10.1002/acr2.11665).

Address correspondence via email to Thomas Hügle, MD, PhD, MA, at Thomas.Hugle@chuv.ch.

Submitted for publication July 26, 2023; accepted in revised form February 9, 2024.

be valuable for clinicians with less experience or without direct access to radiologists. In addition to providing diagnostic support for routine hand radiographs, automated radiographic grading in clinical trials has the potential to save time and enhance research in this field.

Automated machine learning (autoML) or so called no-coding platforms, which allow users to apply machine learning (ML) algorithms, including neural networks, without coding experience, are becoming increasingly popular.[8] These platforms are of particular interest to health care professionals, who have access to clinical data or images but lack coding skills or the ability to collaborate with data scientists, as they allow these professionals to develop their own algorithms.[9] For image classification, autoML platforms enable an end-to-end process, from data upload and augmentation to the training of a CNN and integration into a user interface (eg, a web application).

A few studies have compared the performance of algorithms generated on autoML platforms to the performance of those created by data scientists and found similar results.[10,11] Faes et al explored the feasibility of creating automated deep learning models for medical image classification by health care professionals without coding experience using Google AutoML.[8] Although the models produced by Faes et al showed good performance on internal test sets, the performance on external test sets has room for improvement.[8]

The potential value of user-centric algorithms is broad, including support for health care and administrative workforces in smart hospitals and quality control. It is important to test the user experience of such algorithms via integrated user interfaces or web applications with future target groups. In this study, we demonstrate the full process for the detection and grading of hand OA in a large Swiss arthritis cohort.

MATERIALS AND METHODS

Training, test, and external validation data sets. A total of 13,690 hand radiographs (6,706 of the right hand and 6,984 of the left hand) from 2,863 patients with concomitant RA were extracted from the Swiss Clinical Quality Management in Rheumatic Diseases (SCQM) registry. Patient characteristics are shown in Table 1. As an external test set, 346 DIP joints were extracted, independent of the indication, from 86 hand radiographs of patients aged >55 years but without a history of RA from the radiology department of Centre Hospitalier Universitaire Vaudois (Lausanne University Hospital). Ethical approval from the local committee was obtained for this study.

Radiographic assessment. For the purpose of another study,[12] DIP-OA from the training set was scored by a trained radiology resident according to the modified Kellgren/Lawrence (K/L) score on DIP joints[13] 2 to 5. On each joint, the modified K/L score defined severity as no OA (grade 0), questionable osteophyte(s) and/or joint space narrowing (grade 1), definite small osteophyte(s) and/or mild joint space narrowing (grade 2), moderate osteophyte(s) and/or moderate joint space narrowing (grade 3), or severe osteophytes and/or severe joint space narrowing (grade 4). Erosions and subchondral sclerosis may be present in each class. In the training set, the number of images with a K/L score of 0 was 17,404, the number of images with a K/L score of 1 was 5,187, the number of images with a K/L score of 2 was 9,861, the number of images with a K/L score of 3 was 1,984, and the number of images with a K/L score of 4 was 743.

DIP-OA from the external test set was scored by two other radiologists according to the modified K/L score.[13] The number of images with a K/L score of 0 was 88, the number of images with a K/L score of 1 was 31, the number of images with a K/L score of 2 was 142, the number of images with a K/L score of 3 was 58, and the number of images with a K/L score of 4 was 27 (Table 1). To calculate interobserver variability, 56 images from the external test set were scored by both radiologists, allowing us to calculate a κ score of 0.73.

No-coding platform and user experience. The algorithm was developed by a clinician without experience in coding and ML on Giotto (www.giotto.ai, Learn to Forecast [L2F]), an autoML platform. The clinician was trained for three hours on the use of the platform and basic ML knowledge by a developer of the platform. A written tutorial was available and consulted for further questions. Episodic support by a developer was also needed.

The algorithm was transferred into a user interface (web application). The user experience was analyzed in a questionnaire that was sent to five radiologists and five rheumatologists (Table S1). They were given the choice to either upload their own radiographs of DIP joints to the web application or upload 10 provided precropped images of DIP joints. The corresponding heatmaps were also separately provided.

Table 1. Data set characteristics*

	Mean age (±SD)	Male proportion, %	Female proportion, %	K/L 0, n	K/L 1, n	K/L 2, n	K/L 3, n	K/L 4, n
Training set	60.4 (±10.6)	24	76	17,442	5,120	9,874	1,994	749
Test set	-	-	-	88	31	142	58	27

*K/L, Kellgren/Lawrence score.

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Segmentation model and joint extraction. For this study, two models were trained: a segmentation model used for joint recognition and extraction and a classification model for OA severity scoring. Because the classification task is performed with the whole image provided to the classification model, the image needs to depict only one joint. Joint recognition was done by a segmentation model capable of recognizing each DIP joint in the hand. This model was trained on Giotto by the clinician. DIP joints, proximal interphalangeal joints, metacarpophalangeal joints, and carpometacarpal joints were labeled by the clinician on 519 single hand radiographs. Data augmentation was performed, and the model was trained for 37 epochs. The neural network architecture was a ResNet34. The platform automatically and randomly split the data set into a training set (419 images) and a validation set (100 images). The segmentation algorithm was then used to obtain a mask depicting each joint for each hand radiograph. Python code was written by a developer to allow the DIP joints to be extracted from the masks. Each extracted joint was evaluated by the clinician and was considered correctly extracted if the full joint was represented on the picture without any adjacent joint. All incorrectly extracted joints were discarded. A total of 48,892 (94%) DIP joints were correctly extracted.

Classification model and DIP-OA scoring. The 48,892 scored and extracted DIP joints were used for the training of a classification model to predict the individual K/L score. Among the 48,892 DIP joint images, 10% (4,888 images) were taken out of the data set and used as an internal test set. The 43,973 remaining images were uploaded on the platform. The images in the internal test set came from different patients to those in the training set. Among the images uploaded on the platform, 20% (8,794) were automatically and randomly selected by Giotto as a validation set and were not used to train the algorithm. The classification model was trained on Giotto by the clinician. Data augmentation was performed with the following transformations conducted: rotation, contrast, vertical flip, horizontal flip, brightness, and symmetric warp. The model used was ResNet34, and the number of epochs was 53. When testing the model, the K/L scores 0 and 1 were grouped together and called "no OA."

Heatmap. We performed a heatmap analysis outside the Giotto platform to improve interpretability of the model. Heatmaps allow for the identification of the regions of the image deemed most important by the algorithm for prediction. For a fixed output class, heatmaps identify the regions of the image where the largest gradients of the loss function over the activation function of a chosen inner convolutional layer are located. Thus, the algorithm[14] can be intuitively explained by considering that a change in the network classification decision is mostly due to large gradients: this is a consequence of the stochastic gradient descent[15] algorithm used to train the network. Hence, regions of large gradients would highlight the areas of the image that mostly influence the final classification. Thus, by overlaying the heatmap on the original image, it becomes easier to interpret the model's behavior and identify the specific regions that drive the network's predictions.

Statistics. The model performances were tested in terms of accuracy, sensitivity, and specificity. For the segmentation model, the Dice score and the Mean Intersection Over Union (MIOU) score were automatically calculated by the platform. The quadratic κ coefficient was calculated to evaluate the interobserver variability between the model and the human who scored the data set.

RESULTS

Overall process and handling of the platform. The end-to-end process of the autoML platform from data upload to user interface deployment and evaluation is shown in Figure 1. Because the classification model could only score one joint at a time, two subsequent models were created. First, single hand radiographs were uploaded and their joints were labeled. A segmentation model for DIP joint recognition was trained and used to produce a mask for each radiograph of the data set identifying each joint. This mask was used to extract DIP joints from the data set of single hand radiographs and to upload them on the platform for the training of a classification model. Given the large data set, data upload was performed by the administrator and not by the clinician. They were used as a training set for the generation of a classification model for DIP-OA scoring according to the K/L score. For each model, the platform performed different steps of data augmentation. The platform provided confusion matrixes and accuracy. Heatmaps were produced separately by the administrator (see Heatmaps section). Integration of the model in a web application with the classification model, but not the segmentation model, was done on the platform by the clinician.

Algorithm performances. For the joint segmentation model, the MIOU was 0.79 and the Dice score was 0.88. The loss function across epochs is shown in Figure S1. From the 51,980 scored DIP joints, 48,862 were correctly extracted. The extraction success rate was 94% (see Materials and methods for the metric's definition).

On the internal test set of 4,888 images, the accuracy was 75.5% and the κ score was 0.76 (Table 2). The confusion matrix is shown in Figure 2, and the loss function across epochs is shown in Figure S1. Sensitivity and specificity of the detection of OA were 79% and 86%, respectively.

Tested with an external test set of 346 images, the model showed an accuracy of 66%, and the quadratic κ score was 0.75. The accuracy for each K/L class was as follows: no OA, 79%; K/L class 2, 65%; K/L class 3, 40%; and K/L class 4, 70%

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK

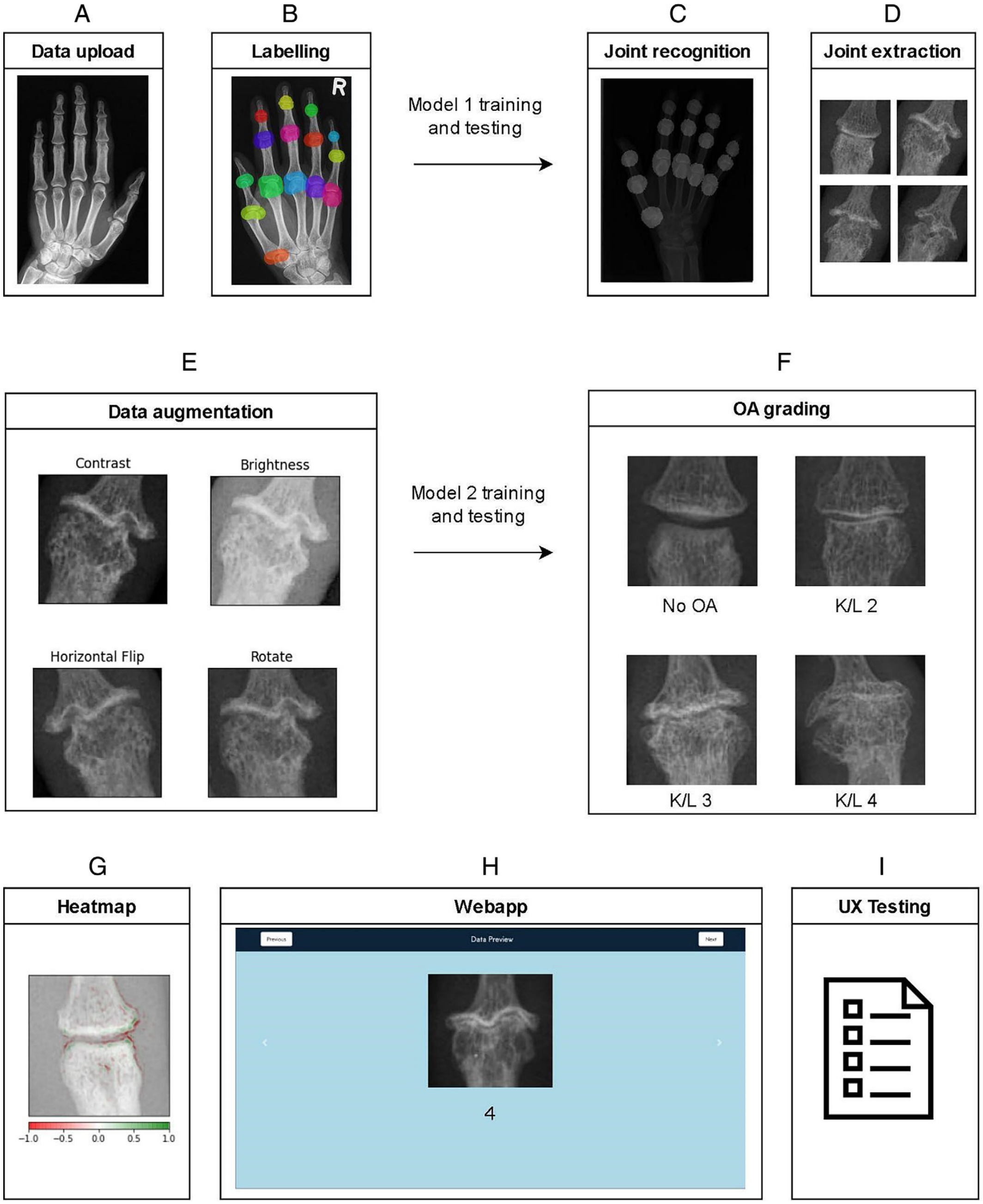

Figure 1. Overview of the autoML procedure from data upload to UX testing via the integrated web application. (A) The data set is uploaded to Giotto following a proper structure indicated by the platform documentation. (B) A segmentation model (model 1) is first created for joint extraction. The task begins by labeling each joint on each hand radiograph (519 hand radiographs). After labeling, the segmentation model is trained. (C) The whole data set (13,690 hand radiographs) is passed to this model, and a mask is obtained for each image. This mask depicts the different joints identified by the model. (D) The mask for each joint is used to extract the DIP joints (48,892 images). (E) The DIP joint images are uploaded to the platform, and a second model is created. This model is a classification model that will classify each DIP joint according to the K/L score. The uploaded data are preprocessed and augmented with different kinds of transformations (eg, vertical flip, rotation, brightness). (F) The model is used as a classifier that allows the grading of DIP-OA on a test set. (G) On request, a heatmap can be generated outside the platform. The heatmap shows the regions of maximum interest for the model's prediction. (H) Once the model's performances are satisfactory, a web application is deployed, allowing anyone to use the model at their own convenience. (I) UX and satisfaction are evaluated through a survey. autoML, automated machine learning; DIP, distal interphalangeal; K/L, Kellgren/Lawrence; OA, osteoarthritis; UX, user experience.

(Table 2). To take into account the unbalanced testing set, in which a majority was K/L class 0, we calculated the mean accuracy, which was 63%. This was done by summing the accuracy for each class and dividing it by the number of classes. Sensitivity and specificity for the detection of OA, regardless of its severity, were 79% and 80%, respectively. The confusion

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. Accuracy*

	Accuracy, %	κ score	No DIP-OA accuracy, %	K/L 2 accuracy, %	K/L 3 accuracy, %	K/L 4 accuracy, %	Mean accuracy, %	Sensitivity for DIP-OA detection, %	Specificity for DIP-OA detection, %
Internal test set	75.5	0.76	86	71	46	67	67.5	79	86
External test set	66	0.75	79	65	40	70	63.5	79	80

*No DIP-OA was defined as K/L 0 and 1. DIP-OA, distal interphalangeal osteoarthritis; K/L, Kellgren/Lawrence score.

matrix is shown in Figure 2. Another model with a balanced training set (approximately 850 images in each class randomly selected from the original data set) was also developed, and its results were not better than the results of the unbalanced model (Table S1 and Figure S2).

Heatmap. Heatmaps showing the regions of highest importance for class prediction were obtained (Figure 3). They were used as a quality control to assess correct structure recognition by the model (osteophytes, joint space narrowing). Images correctly classified as no OA showed a pattern along the conserved joint space. In images classified as grade 2, the region of interest was at the osteophytes and focal joint space narrowing. In images classified as grade 4, osteophytes, joint space narrowing, and central erosion were correctly identified by the model. A totally fused joint was misclassified as having grade 2 OA. The model appears to have mistaken an enthesophyte as an osteophyte (Figure S3).

User experience testing. The user experience of the second model (DIP-OA classification) was tested in focus groups with five radiologists and five rheumatologists. Among the five radiologists, one found an algorithm for automated scoring of DIP-OA moderately useful, and the rest found it of little use because all of them were familiar with the different stages of hand OA. In contrast, all five radiologists considered the heatmaps to be very helpful for understanding and trusting in the model. Among the five rheumatologists, three found the web application very useful and two found it moderately useful to detect and grade DIP-OA because they were relatively unfamiliar with its radiographic progression. All rheumatologists appreciated the algorithm as a potential tool to differentiate OA from inflammatory diseases such as psoriatic arthritis (mean 9.2, SD 1.1) and as moderately important for documenting grading of hand OA over time (mean 6.2, SD 1). Four of five rheumatologists found heatmaps to be useful for understanding and trusting in the model, but only two of five believed that the heatmap would be of added value in the web application (Table S2).

DISCUSSION

This study introduces a robust algorithm for detecting and scoring DIP-OA in hands. Specifically, we outline an end-to-end process using an autoML platform to generate a CNN model

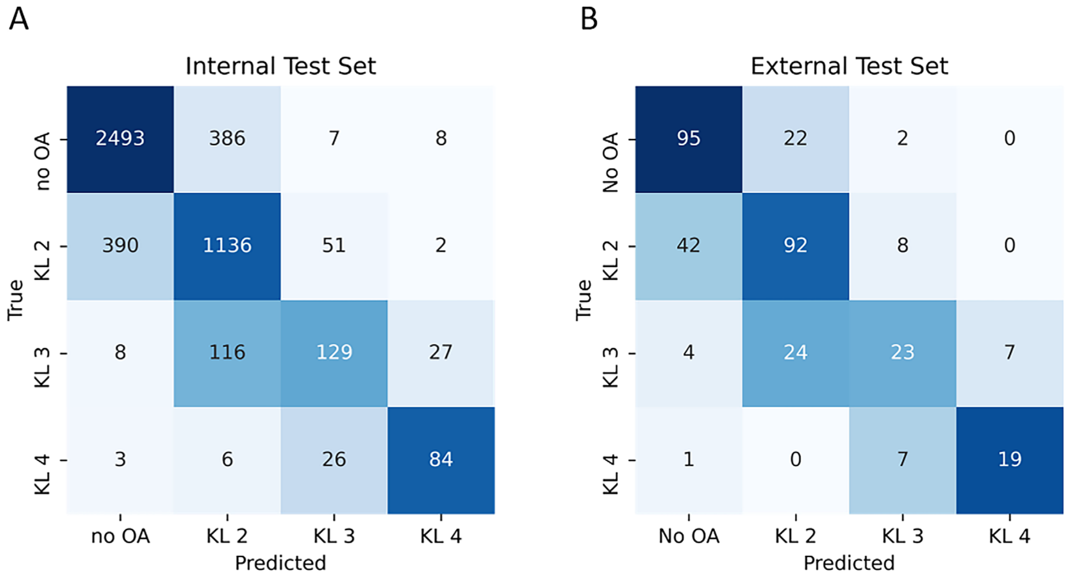

Figure 2. Confusion matrix for the (A) internal test set and (B) external test set. The confusion matrix displays the comparison of the model's predictions and the truth (graded by the radiologists) for the images of the external test set. On the x-axis for each class, the corresponding number of predicted images is reported. On the y-axis for each class, the corresponding number of images according to the human grader (true) is reported. The confusion matrix allows us to evaluate the model's performances for each class. KL, Kellgren/Lawrence score.

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

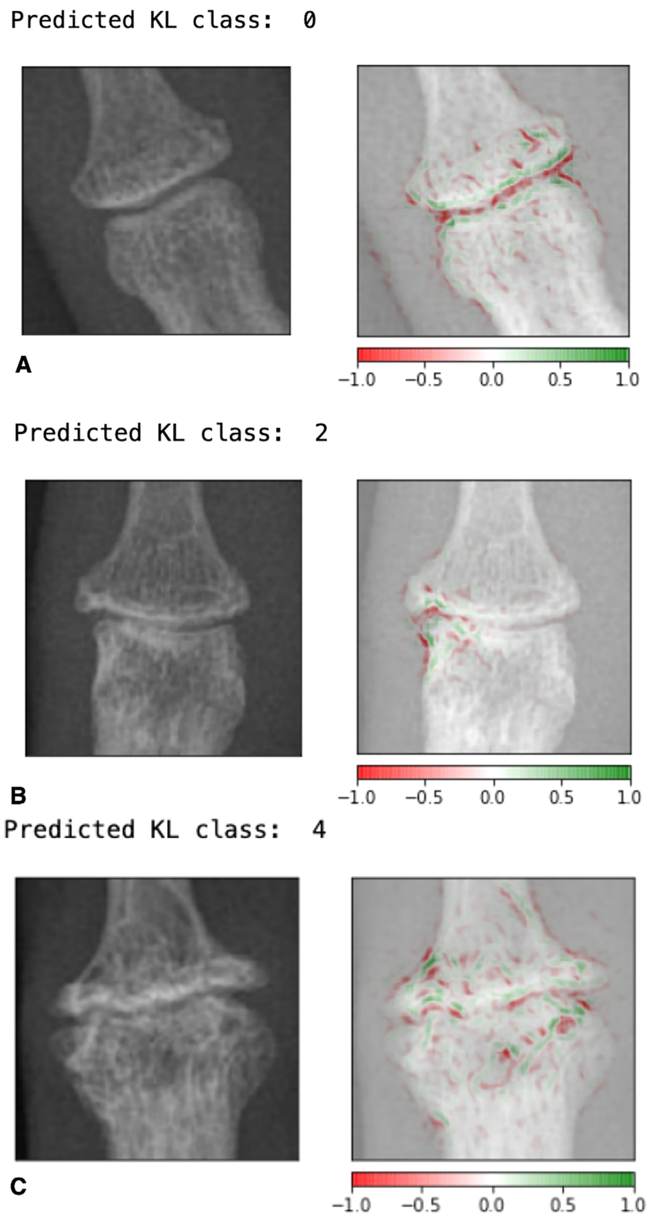

Figure 3. Heatmaps of different joints from the external test set. The delineated portions of the images represent parts of the image with a higher gradient and thus are of higher interest for the prediction. (A) A normal joint. The most important parts for the prediction were located at the joint space and at the joint margin, which correspond to the features used for osteoarthritis grading by humans. The model achieved a correct prediction of KL class 0 (B) Mild OA. The gradients were higher at the left joint margin and joint space. The image depicts an osteophyte and focal joint space narrowing at this place. The model correctly classified it as KL class 2. (C) Severe OA. The model identified severe joint space narrowing and severe osteophytes, which led to the correct prediction of KL class 4. KL, Kellgren/Lawrence.

and tested the user interface in two target user groups. With a data set comprising more than 13,000 hand radiographs, the model demonstrated good performance in detecting hand OA, achieving a sensitivity and specificity of approximately 80%. The accuracy for grading OA severity based on the K/L score varied, ranging from 40% to 86% for different grades and revealing the algorithm's limitations, which may also reflect challenges in human grading of hand OA radiographs. The considerable interobserver variability of the modified hand OA K/L score (κ from 0.55 to 0.8) outlines the difficulty of grading OA.[13,16,17]

Rheumatologists found this tool more useful than radiologists, which probably is inherent in the nature of the matter and may also be due to the fact that rheumatologists have prescribed and interpreted fewer hand radiographs since the advent of ultrasound. All users found accompanying heatmaps useful, highlighting regions of interest in individual images. Heatmaps, as attractive clinical decision support tools, have been recognized by others in various medical contexts, such as in the evaluation of pathology slides.[18] As illustrated in the heatmaps in Figure 3, each predicted class exhibits distinctive features corresponding to the key anatomic characteristics of each K/L score (a preserved joint space in grade 0 or 1, small osteophytes in grade 2, and significant joint space narrowing with a prominent osteophyte in grade 4). In misclassified images, heatmaps can indicate nonanatomic discriminant features. Illustrated in Figure S3A, our model misidentified a fused joint as a K/L score of 2, mistakenly interpreting an enthesophyte as an osteophyte and the epiphyseal line as joint space because the model lacked training to recognize ankylosis. However, heatmaps are not designed to elucidate the rationale behind the prediction of one grade over another. In a different instance (Figure S3B), the model accurately identified a conserved joint space and thus predicted a K/L score of 0. Conversely, the radiologist graded this image as mild OA (K/L score of 2) because of focal joint space narrowing. In this example, the heatmap was not able to explain the misclassification.

The model itself achieved a similar performance to that of radiologists reported in the literature.[13] Compared to previous knee OA CNN prediction models, our model has a similar mean accuracy[19] of 67%. This is assuring because the radiographic progression in DIP-OA is more complex than that in knee OA, notably with a nonlinear dynamic of the joint space with repair phases and erosions. Additionally, compared to other autoML platforms, our model performed similarly.[8]

An implication of this work is that AI support as an adjunct will likely appear for hand OA because it is already the case for knee OA and other indications. We postulate that heatmaps are a candidate to be integrated into the clinical workflow, accessible via the electronic medical record (EMR). The future use of automated AI platforms is more controversial. Obviously, algorithms generated by autoML such as this one cannot be simply implemented in the routine because they require medical device certification. In our example, two algorithms needed to be connected to extract the correct joints and then to classify the OA grade. For the large number of images used here, the drag and drop function did not work. In other words, the use of autoML platforms still requires training or at least support from data scientists. In terms of research, however, autoML platforms give clinicians and

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

researchers new options to better analyze clinical or preclinical data. This refers especially to large data sets such as EMR systems, registries such as the one used in this study, and experimental preclinical data sources such as histologic slides. As an example, we would like to mention ML algorithms created from deidentified EMR databases, such as Epic, which uses the data of more than 180 million patients in their Cosmos program. It comes as no surprise that services have been established to reply to specific clinical queries within the clinical workflow.[20]

Our study is subject to several limitations. Accuracy may not be the most suitable metric indicator in this context. Instead, performance could be more accurately conveyed through the quadratic kappa score, which assesses the variability between two different observers, such as the radiologist and the algorithm, taking into consideration the continuous nature of the task. Given the substantial interobserver variability inherent in the K/L score, our model's performance is deemed satisfactory. Concerning the grading of hand OA, the model exhibited lower performance. Notably, the accuracy in predicting DIP-OA grade 3 posed a significant challenge, impacting the reliability of the model for use in clinical trials. This difficulty may be attributed to the limited representation of this class in the training set, suggesting potential improvement in subsequent studies. Furthermore, the underrepresentation of severe DIP-OA in our model could introduce biases during training. Interestingly, attempts to rectify this imbalance through a balanced data set, with equal distribution among the classes, did not result in a significant enhancement in recognizing severe OA (see Table S1 and Figure S2). The fact that we trained this DIP-OA model in patients with an RA background is another limitation. Of note, RA normally does not affect the DIP joints, and in a previous study of the same data set we demonstrated that RA disease activity was not associated with radiographic DIP-OA progression.[12] In the external test set without concomitant RA, the model was effective despite its performance being slightly better in the internal data set. This could be explained by the fact that the training set and internal test set of images were labeled by one single radiologist, whereas the external validation images were labeled by several radiologists. This difference could be attributed to interobserver variability and may represent another limitation. Finally, we only trained the model on a single autoML platform and tested the user experience properties only in the provided web application. Potentially, alternative platforms with more intuitive web applications would have convinced radiologists about the use of hand OA prediction.

Nonetheless, the simplicity of use of autoML platforms makes ML models accessible to more clinicians and scientists. There are already such applications in which concrete clinical questions are entered and automated reports and publishable graphics via a dashboard are received.[21] Care must be taken to ensure quality of the generated models and robust assessment of their performance. To be reliable, an ML model must have been trained on quality and diverse data, and its generalizability must have been tested by an external test set.[22] Models trained on dubious data in term of quality or size can produce extremely good results when tested internally, but they may not fare as well when applied to real-world data.[8] For this reason, guidelines such as SPIRIT-AI and DECIDE-AI have been published to guide clinical trials on ML models.[23,24] The simplicity of use of these platforms is as much an asset as it is a danger, and their clinical use (eg, in using EMR data) must follow regulation. Most of all, the usefulness of the task of the algorithm should be questioned. In our case, we see a clear value of automated hand OA detection to answer quicker research questions (eg, in larger radiographic data sets, such as the Osteoarthritis Initiative) and to correlate it with other variables.

In conclusion, autoML platforms are an innovative tool that can make data science and ML more accessible to develop sustainable user-centric algorithms, such as the automated detection of radiographic DIP-OA. However, we believe that those platforms have the responsibility to raise awareness on how to develop trustable models.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr Hügle had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Caratsch, Hügle.

Acquisition of data. Lechtenboehmer, Oung, Zanchi, Aleman.

Analysis and interpretation of data. Caratsch, Caorsi, Walker, Omoumi, Hügle.

REFERENCES

1. Hügle M, Omoumi P, van Laar JM, et al. Applied machine learning and artificial intelligence in rheumatology. Rheumatol Adv Pract 2020;4(1): rkaa005.
2. Radiobotics Nabs 510(k) clearance for knee osteoarthritis software. FDAnews. September 2, 2021. Accessed March 12, 2023. https://www.fdanews.com/articles/204230-radiobotics-nabs-510k-clearance-for-knee-osteoarthritis-software
3. Hirano T, Nishide M, Nonaka N, et al. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. Rheumatol Adv Pract 2019;3(2):rkz047.
4. Binvignat M, Pedoia V, Butte AJ, et al. Use of machine learning in osteoarthritis research: a systematic literature review. RMD Open 2022; 8(1):e001998.
5. Kloppenburg M, Kroon FP, Blanco FJ, et al. 2018 update of the EULAR recommendations for the management of hand osteoarthritis. Ann Rheum Dis 2019;78(1):16–24.
6. Üreten K, Maraş HH. Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods. J Digit Imaging 2022;35(2):193–199.
7. Verbruggen G, Veys EM. Erosive and non-erosive hand osteoarthritis. Use and limitations of two scoring systems. Osteoarthritis Cartilage 2000;8(suppl A):S45–S54.
8. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

coding experience: a feasibility study. Lancet Digit Health 2019;1(5): e232–e242.

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

10. Wan KW, Wong CH, Ip HF, et al. Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. Quant Imaging Med Surg 2021;11(4):1381–1393.

11. Musigmann M, Akkurt BH, Krähling H, et al. Testing the applicability and performance of Auto ML for potential applications in diagnostic neuroradiology. Sci Rep 2022;12(1):13648.

12. Lechtenboehmer CA, Jaeger VK, Kyburz D, et al. Influence of disease activity in rheumatoid arthritis on radiographic progression of concomitant interphalangeal joint osteoarthritis. Arthritis Rheumatol 2019;71(1):43–49.

13. Zhang Y, Niu J, Kelly-Hayes M, et al. Prevalence of symptomatic hand osteoarthritis and its impact on functional status among the elderly: the Framingham Study. Am J Epidemiol 2002;156(11):1021–1027.

14. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. Paper presented at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22–29, 2017; Venice, Italy. doi:https://doi.org/10.1109/ICCV.2017.74

15. Ruder S. An overview of the gradient descent optimization algorithms. arXiv Preprint posted online June 15, 2017. doi:10.48550/arXiv.1609.04747

16. Haugen IK, Slatkowsky-Christensen B, Bøyesen P, et al. Cross-sectional and longitudinal associations between radiographic features and measures of pain and physical function in hand osteoarthritis. Osteoarthritis Cartilage 2013;21(9):1191–1198.

17. Visser AW, Bøyesen P, Haugen IK, et al. Radiographic scoring methods in hand osteoarthritis–a systematic literature search and descriptive review. Osteoarthritis Cartilage 2014;22(10):1710–1723.

18. Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digit Health 2020;2(8):e407–e416.

19. Tiulpin A, Thevenot J, Rahtu E, et al. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep. 2018;8(1):1727.

20. Tarabichi Y, Frees A, Honeywell S, et al; The Cosmos Collaborative. The Cosmos Collaborative: a vendor-facilitated electronic health record data aggregation platform. ACI Open 2021;5(1):e36–e46.

21. Callahan A, Gombar S, Cahan EM, et al. Using aggregate patient data at the bedside via an on-demand consultation service. NEJM Catal Innov Care Deliv 2021;2(10). doi:https://doi.org/10.1056/CAT.21.0224

22. Volovici V, Syn NL, Ercole A, et al. Steps to avoid overuse and misuse of machine learning in clinical research. Nat Med 2022;28(10):1996–1999.

23. Cruz Rivera S, Liu X, Chan AW, et al; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2020;2(10):e549–e560.

24. Vasey B, Nagendran M, Campbell B, et al; DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022;28(5):924–933.

25785745, 2024, 6, Downloaded from https://acrjournals.onlinelibrary.wiley.com/doi/10.1002/acr2.11665 by Sissa Scuola Internazionale, Wiley Online Library on [23/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RIGHTSLINK
Copyright Clearance Center