doi:10.1093/rap/rkz047

RHEUMATOLOGY ADVANCES IN PRACTICE

Original article

Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis

Toru Hirano [1], Masayuki Nishide[1], Naoki Nonaka[2], Jun Seita[2], Kosuke Ebina[3], Kazuhiro Sakurada[2] and Atsushi Kumanogoh[1]

Abstract

Objective The purpose of this research was to develop a deep-learning model to assess radiographic finger joint destruction in RA.

Methods The model comprises two steps: a joint-detection step and a joint-evaluation step. Among 216 radiographs of 108 patients with RA, 186 radiographs were assigned to the training/validation dataset and 30 to the test dataset. In the training/validation dataset, images of PIP joints, the IP joint of the thumb or MCP joints were manually clipped and scored for joint space narrowing (JSN) and bone erosion by clinicians, and then these images were augmented. As a result, 11 160 images were used to train and validate a deep convolutional neural network for joint evaluation. Three thousand seven hundred and twenty selected images were used to train machine learning for joint detection. These steps were combined as the assessment model for radiographic finger joint destruction. Performance of the model was examined using the test dataset, which was not included in the training/validation process, by comparing the scores assigned by the model and clinicians.

Results The model detected PIP joints, the IP joint of the thumb and MCP joints with a sensitivity of 95.3% and assigned scores for JSN and erosion. Accuracy (percentage of exact agreement) reached 49.3–65.4% for JSN and 70.6–74.1% for erosion. The correlation coefficient between scores by the model and clinicians per image was 0.72–0.88 for JSN and 0.54–0.75 for erosion.

Conclusion Image processing with the trained convolutional neural network model is promising to assess radiographs in RA.

Key words: rheumatoid arthritis, joint destruction, artificial intelligence

CLINICAL SCIENCE

Key messages

- Convolutional neural network-based deep learning can be applied to develop a model for assessing hand radiographs.
- The model assesses joint space narrowing and bone erosion of the fingers of RA.
- This artificial intelligence technology will lead to more extensive and detailed evaluation of joints in future.

[1]Department of Respiratory Medicine and Clinical Immunology, Internal Medicine, Graduate School of Medicine, Osaka University, Suita, Osaka, [2]Medical Sciences Innovation Hub Program, RIKEN, Yokohama, Kanagawa and [3]Department of Musculoskeletal Regenerative Medicine, Graduate School of Medicine, Osaka University, Suita, Osaka, Japan

Submitted 3 September 2019; accepted 4 November 2019

Correspondence to: Toru Hirano, Department of Respiratory Medicine and Clinical Immunology, Internal Medicine, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka 565-0871, Japan. E-mail: thirano@imed3.med.osaka-u.ac.jp

© The Author(s) 2019. Published by Oxford University Press on behalf of the British Society for Rheumatology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Artificial intelligence (AI) is used effectively in a wide range of fields, including autonomous vehicles, translation, speech recognition, image processing, natural language processing, art and medicine [1]. In medicine, processing of histopathological images or radiographs, mining of genomic data, screening for molecular targets and analysis of clinical big data are considered suitable tasks for deep learning, in which multiple layers of neuron-like nodes mimic how human brains analyse information [2]. AI trained with deep learning, such as a convolutional neural network (CNN), which introduces robustness to variations of images, can deliver outstanding performance in classifying various images into significant categories [3]. Examples include the grading of diabetic retinopathy [4], classification of skin cancer [5, 6], prediction of lung cancer mutations [7] and classification of interstitial lung disease [8].

The management of RA has progressed dramatically over the past several decades. With the use of numerous efficacious drugs, more than half of patients have achieved low disease activity or clinical remission, which includes less joint pain, less joint swelling, lower serum inflammatory markers and better global health based on patients' self-assessment. Disease activity and other factors, such as progression of structural joint damage, are considered to make treatment decisions [9]. For evaluating structural joint damage, radiography has been the gold standard.

Radiographic classification of RA or systemic arthritis was first proposed by Steinbrocker *et al*. [10]. Several methods for scoring radiographic joint damage in RA were proposed by Kellgren & Bier [11], Sharp *et al*. [12, 13], Larsen *et al*. [14] or Genant [15]. Later, the Sharp/van der Heijde method for scoring radiographs of hands and feet was developed [16]. This method has been widely used, especially in clinical studies. However, in clinical settings, this method is not commonly used because it requires a high level of skill, and the differences between examiners are considerable.

In the present research, we attempt to develop a model for scoring of radiographic finger joint destruction in RA. Our model comprises two steps, as shown in Fig. 1A. The first step is a detection of joints by machine learning (cascade classifier using Haar-like features). The second step is a scoring of joint destruction by deep learning (CNN), which comprises convolutional layers, pooling layers and fully connected layers. CNN processes the input image as two-dimentional matrix data and gives output as numerical values or probability of categories. CNN is currently considered to be the most efficient algorithm for image processing. We examine the performance of our model by comparing scores assigned by the model and those assigned by rheumatologists using radiographs that were not included in the CNN training/validation process.

Methods

Patients and images

Digital anterior–posterior radiographs of front bilateral hands of 108 patients with RA were collected retrospectively. Patients were diagnosed with RA according to the 1987 Rheumatoid Arthritis Classification by the ARA [17] or the 2010 ACR–EULAR Classification Criteria [18]. All patients were treated at Osaka University Hospital and were enrolled in the institute's cohort of RA patients. Clinical information, such as sex, age, disease duration and clinical laboratory values, was collected from the medical charts. This research was approved by the ethics committee of Osaka University Hospital and was conducted in accordance with the Declaration of Helsinki.

Among 216 radiographs of the 108 patients, we used 186 radiographs of 93 patients for training/validation, and 30 radiographs of 15 patients for testing of the trained model (Fig. 1B). From the 186 radiographs, areas of PIP joints, the IP joint of the thumb or MCP joints were manually clipped with the use of a graphical software package, and 1860 clipped images were generated. The images were grey scaled, with the resolution ranging from 40×40 pixels to 80×80 pixels. The degree of joint destruction was scored by consensus between two rheumatologists, with 10 or 15 years of experience in rheumatology, according to the Sharp/van der Heijde method [19]. The scores consisted of the joint space narrowing (JSN) score and the bone erosion score. Briefly, the JSN score is defined as follows: 0, no JSN; 1, focal or doubtful; 2, generalized, >50% of the original joint space left; 3, generalized, <50% of the original joint space left or subluxation; and 4, bony ankylosis or complete luxation. The erosion score is defined as follows: 0, no erosion; 1, discrete; 2, larger, <50% of the joint surface; 3, extending over the middle of the bone; and 5, complete collapse. The erosion score of a single joint is calculated as the sum of each score in the joint, with a maximal score of 5.

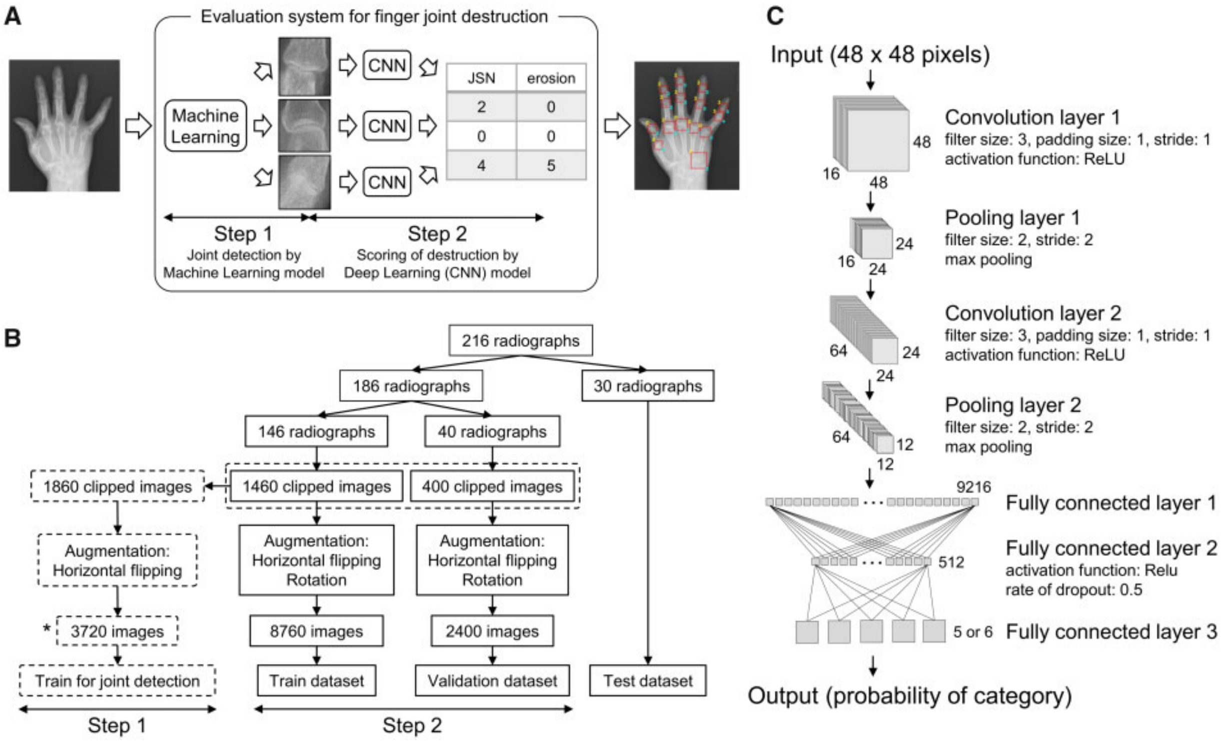
Detection of joints by machine learning

The first step of machine learning was to detect the finger joints. The finger joints were detected by a cascade classifier using Haar-like features [20]. The classifier was trained to detect finger joints such as PIP, IP and MCP joints. The 1860 clipped images from 186 radiographs were augmented by horizontal flipping, and a total of 3720 images was used to train the classifier (Fig. 1B). For training the classifier and applying joint detection, Open Source Computer Vision Library, Open-CV (v.3.4; Intel Corporation, Santa Clara, CA, USA) was used.

Scoring of joint destruction by CNN

The second step of machine learning was to assign a JSN score and an erosion score to each joint detected in the previous step. The collection of the original radiographs were split into three sets, namely, training,

Fig. 1 Flow of machine learning


A
Evaluation system for finger joint destruction
Machine Learning
CNN
CNN
CNN
JSN
erosion
2
0
0
0
4
5
Step 1
Joint detection by Machine Learning model
Step 2
Scoring of destruction by Deep Learning (CNN) model
B
216 radiographs
186 radiographs
30 radiographs
146 radiographs
40 radiographs
1860 clipped images
1460 clipped images
400 clipped images
Augmentation: Horizontal flipping
Augmentation: Horizontal flipping Rotation
Augmentation: Horizontal flipping Rotation
* 3720 images
8760 images
2400 images
Train for joint detection
Train dataset
Validation dataset
Test dataset
Step 1
Step 2
C
Input (48 x 48 pixels)
Convolution layer 1
filter size: 3, padding size: 1, stride: 1
activation function: ReLU
16
48
48
Pooling layer 1
filter size: 2, stride: 2
max pooling
16
24
24
Convolution layer 2
filter size: 3, padding size: 1, stride: 1
activation function: ReLU
64
24
24
Pooling layer 2
filter size: 2, stride: 2
max pooling
64
12
12
9216
Fully connected layer 1
Fully connected layer 2
activation function: Relu
rate of dropout: 0.5
512
5 or 6
Fully connected layer 3
Output (probability of category)

(**A**) The first step of the machine learning is a detection of finger joints, and the second step is a scoring of joint destruction. These steps are combined as the assessment model for radiographic finger joint destruction. (**B**) The first step used 3720 images for machine learning (*). The second step used 8760 images derived from 146 radiographs for training (train dataset) the convolutional neural network (CNN), and 2400 images derived from 40 radiographs for validation during the training process (validation dataset). Thirty radiographs were used for testing the performance of the model (test dataset). (**C**) The network of the CNN consists of two convolution layers, two pooling layers and three fully connected layers.

validation and testing datasets, following the standard practice in machine learning (Fig. 1B). Each dataset contained 146, 40 and 30 radiographs, respectively. Radiographs contained in the training dataset were used to tune the parameters of the CNN, and radiographs contained in the validation dataset were used to monitor the performance of the CNN model during the training process. After the training process, radiographs in the testing dataset were used to evaluate the trained CNN model by comparing scores assigned by the model and by clinicians. In the training process, radiographs in the training and validation datasets were augmented by horizontal flipping and/or rotation (+10 or −10°). As a result, we obtained a total of 8760 and 2400 images of PIP, IP or MCP joints for the training and validation dataset, respectively. Subsequently, the obtained images were resized to 48×48 pixels and offered to the CNN model. The CNN model comprises two convolution layers [filter size: 3; padding size: 1; stride: 1; activation function: rectified linear unit (ReLU)], two pooling layers (filter size: 2; stride: 2, maximal pooling) and three fully connected layers with one hidden layer (512 nodes; activation function: ReLU; rate of dropout: 0.5) (Fig. 1C). The loss function was set to softmax cross entropy, and the optimization algorithm was set to adaptive moment estimation (Adam) [21]. The batch size for the training was set to 512. Batch normalization was introduced in the CNN for the erosion score based on the preliminary experiment [22]. Output was given as the probability of each JSN class or each erosion class, and the class with the highest probability was determined. The Open Source Library for Neural Networks, Chainer (v.5.1; Preferred Networks, Tokyo, Japan) was used for implementation of the CNN model [23].

Testing of the model

To test the performance of the trained model, we assessed the consistency of judgements between the model and two clinicians. One of the two clinicians (Clinician 2) was officially trained for scoring of joint destruction. Thirty radiographs in the test dataset were used (Fig. 1B). The numbers of each JSN class or erosion class assigned by the model or clinicians were counted, and the distributions of scores were compared. The percentage of exact agreement (PEA), which is

Table 1 Characteristics of the patients

Characteristic	Total	Train/validation	Test
n	108	93	15
Sex, female/male	90/18	77/16	13/2
Age, years	64.9 (53.5, 72.6)	64.9 (53.4, 72.6)	64.2 (56.8, 76.0)
Disease duration, years	12.2 (6.4, 17.6)	12.3 (6.8, 18.6)	9.4 (0.7, 14.1)
Class I/II/III/IV	39/56/13/0	35/46/12/0	4/10/1/0
Stage I/II/III/IV	28/19/29/32	24/16/26/27	4/3/3/5
ACPA positive, *n* (%)	73 (67.6)	63 (67.7)	10 (66.7)
Number of radiographs	216	186	30

Values of age and disease duration are given as the median and interquartile range. Other values are numbers in each category.

identical to accuracy, and the percentage of close agreement (PCA), which is within 1.0 score difference at the joint level, were assessed. Sensitivity and specificity (score zero *vs* at least one) were also assessed. The total score of a radiograph for JSN or erosion was calculated as the sum of each JSN score or erosion score of PIP, IP and MCP joints. Correlations between total scores assigned by the model and those by clinicians were assessed using Pearson's correlation coefficients.

Results

Patients and images

The characteristics of the patients are shown in Table 1. Among the patients, 90 (83.3%) were female. The median and the interquartile range of age were 64.9 (53.5, 72.6) years old, and those of disease duration were 12.2 (6.4, 17.6) years. All participants were diagnosed with RA. Seropositivity of ACPA was 67.6%. In the training/validation dataset, the distribution of the 1860 clipped images by joint was 744 for PIP joint, 186 for IP joint and 930 for MCP joint. The JSN score and the erosion score assigned by clinicians are summarized in Table 2. Scores for intercarpal joints were not summarized because the model trained by the machine learning could not detect many of them.

Detection of joints

Fig. 2 shows representative images processed by the model, which detected finger joints and then assigned scores of joint destruction. Finger joints, such as PIP, IP and MCP joints, were identified as red rectangles by the model. Fig. 2A shows the whole hand image processed by the model. In this image, four PIP joints, one IP joint and five MCP joints were correctly detected. The DIP joints, some intercarpal and wrist joints were also detected, but many intercarpal joints were not identified correctly.

Scoring of joint destruction

In Fig. 2, the number at the upper left corner of the rectangle indicates the JSN score assigned by the model (yellow letter) and that at the lower right corner indicates the erosion score assigned by the model (blue letter). In Fig. 2B, an enlarged image shows joints with JSN score 0, 2, 3 or 4 and erosion score 0, 4 or 5. In Fig. 2C, another enlarged image shows joints with JSN score 0, 2 or 4 and erosion score 0, 3 or 5. The accuracy (PEA) of scoring during the training/validation process of the CNN increased continuously with epoch, which is the number of repetitions of the training (Fig. 3A for JSN and Fig. 3C for erosion). The loss, which is the discrepancy between the score assigned by the CNN and the score determined by clinicians, decreased with epoch (Fig. 3B for JSN and Fig. 3D for erosion). The training of the CNN was stopped at epoch 40 for JSN and at epoch 110 for erosion, when the values of loss were minimum, respectively. In the validation dataset, the accuracy (PEA) of the JSN score reached 60.6% (Fig. 3a, blue line) and that of erosion score reached 72.6% (Fig. 3C, blue line).

Table 2 Scoring of joint destruction on training/validation dataset

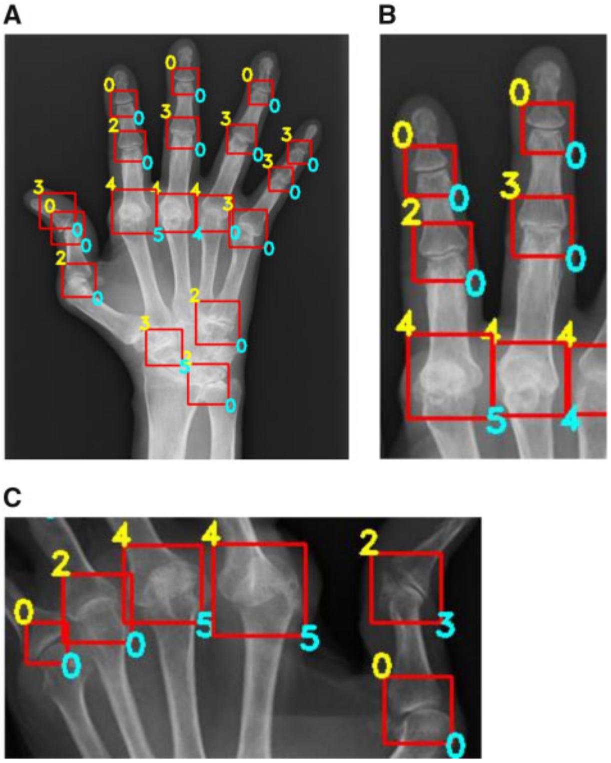
Score	JSN score		Erosion score	
	PIP/IP	MCP	PIP/IP	MCP
0	128	510	644	761
1	58	28	74	22
2	356	184	93	33
3	326	127	28	28
4	62	81	22	15
5	N.D.	N.D.	69	71

IP: IP joint of the thumb; JSN: joint space narrowing; N.D.: not defined.

Testing of the model

The rate of joint detection by the trained model reached 95.3% (286/300). In detail, 98.3% of PIP (118/120), 86.7% of IP (26/30) and 94.0% of MCP joints (142/150) were correctly detected by the model. Joint detection failed in two PIP joints, four IP joints and eight MCP

FIG. 2 A representative image processed by the model



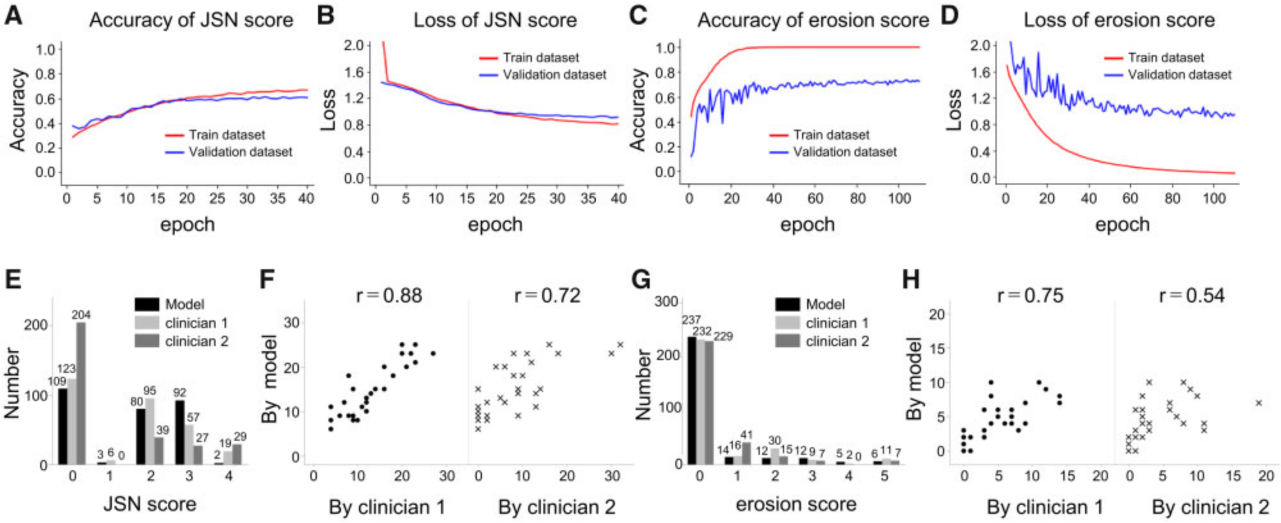
(**A**) A whole hand image processed by the model. The red rectangle indicates joints, such as PIP, IP or MCP. The number at the upper left in the rectangle indicates the joint space narrowing (JSN) score (yellow letter) and that at the lower right indicates erosion score (blue letter). (**B**) An enlarged image shows the joints with JSN score 0, 2, 3 or 4 and those with erosion score 0, 4 or 5. (**C**) Another enlarged image shows the joints with JSN score 0, 2 or 4 and those with erosion score 0, 3 or 5. IP: IP joint of the thumb.

joints. These areas often contained severely impaired bone alignment or luxation. The distributions of scores by the model and clinicians are shown in Fig. 3E for JSN and Fig. 3G for erosion. The consistency of scores by the model and clinicians are summarized in Table 3. PEA (accuracy) between the model and two clinicians was 49.3–65.4% for JSN and 70.6–74.1% for erosion. The percentage of close agreement was 64.0–85.3% for JSN and 84.3% for erosion. Sensitivity and specificity (score zero *vs* at least one) were 88.0–94.2% and 52.0–74.8% for JSN, and 34.8–42.4% and 88.2–89.4% for erosion (Supplementary Tables S1 and S2, available at *Rheumatology Advances in Practice* online). Scatter plots of total score per radiograph are shown in Fig. 3F for JSN and Fig. 3H for erosion. The correlation coefficients between scores by the model and two clinicians were 0.72–0.88 for JSN and 0.54–0.75 for erosion.

Discussion

In this research, we demonstrated how a deep-learning model was trained in order to assess radiographic finger joint destruction in RA. Disease durations of the patients were relatively long (median 12.2 years), and 67.6% of the patients were seropositive for ACPA, which is a strong predictor for radiographic progression in RA [24–26]. Thus, the patients enrolled in this study were predisposed to have joint destruction. Finger joints, such as PIP, IP and MCP joints, were correctly detected by the model with a sensitivity of 95.3%. Intercarpal joints tended to be ignored by the model, probably because images of intercarpal joints were not offered to the machine learning, and the structures of these areas were too complex for the current model to detect. In addition, some joints with severely impaired alignments or luxation were ignored. PEA (accuracy) of the JSN score reached 49.3–65.4%, and that of the erosion score reached 70.6–74.1%. PCA of the JSN score reached 64.0–85.3%, and that of the erosion score reached 84.3%. The percentage of agreement of the JSN score for PIP/IP joints was obviously low (Table 3). Given that machine learning was conducted using images of PIP, IP and MCP joints together, it might be difficult for the model to judge the differences between joints. As shown in Fig. 3G, the distribution of erosion score seems comparable. However, as shown in Fig. 3E, the model and Clinician 1 judged too much for score 0, 2 and 3 compared with Clinician 2. The correlation coefficients between scores by the model and two clinicians were

Fig. 3 Test of the model



(**A**, **B**) The accuracy, identical to the percentage of exact agreement (PEA), and the loss of joint space narrowing (JSN) score during the process for training dataset (red line) and validation dataset (blue line). (**C**, **D**) The accuracy (PEA) and the loss of erosion score for training dataset (red line) and validation dataset (blue line). (**E**) Distribution of the JSN score assigned by the model (black bar) and by clinicians (light and dark grey bars). (**F**) Correlation of JSN score between the model and clinicians. (**G**) Distribution of erosion score assigned by the model (black bar) and by clinicians (light and dark grey bars). (**H**) Correlation of erosion score between the model and clinicians.

Table 3 Consistency of scores by the model and clinicians

Evaluator	Index	Total (%)	PIP/IP (%)	MCP (%)
For JSN				
Model *vs* Clinician 1	PEA	65.4	58.3	72.5
	PCA	85.3	84.0	86.6
Model *vs* Clinician 2	PEA	49.3	24.3	74.6
	PCA	64.0	43.1	85.2
Clinician 1 *vs* 2	PEA	55.5	36.7	74.5
	PCA	67.6	52.7	82.6
For erosion				
Model *vs* Clinician 1	PEA	74.1	66.0	82.4
	PCA	84.3	81.9	86.6
Model *vs* Clinician 2	PEA	70.6	65.2	76.1
	PCA	84.3	81.3	87.3
Clinician 1 *vs* Clinician 2	PEA	70.6	66.0	75.2
	PCA	88.0	88.7	87.2

A total of 286 joints were assessed. Fourteen joints were not identified by the model. PEA is the percentage of exact agreement, and PCA is the ratio of close agreement (within 1.0 score difference) among evaluators.

0.72–0.88 for the JSN score and 0.54–0.75 for the erosion score. In a previous report, correlation coefficients between readings of multiple observers, who were radiologists or rheumatologists, were 0.585–0.947 for the JSN score and 0.529–0.962 for the erosion score [13]. Sensitivity and specificity were 88.0–94.2% and 52.0–74.8% for JSN, and 34.8–42.4% and 88.2–89.4% for erosion. From these results, the most problematic thing in our model was underdiagnosis of erosions and overdiagnosis of JSN.

Thus, image processing techniques for hand radiographs using the CNN might be used in the evaluation of joint destruction in RA. Assessment by the model takes <1 s per image (average 0.63 s). This is obviously faster than the time required by humans to make an assessment. Although echography and MRI examinations are increasingly used for the assessment of joint damage, radiographs retain a unique value by providing a comprehensive or panoramic view of the joints. Automated assessment of radiographs with a deep-learning algorithm would be of great value in many clinical situations. Moreover, novel radiographic findings about joint destruction might be discovered. Recently, many studies using deep learning or CNN for assessing joints or bones have been reported. These include diagnosis of hip OA [27], bone age assessment [28, 29], fracture detection [30] and assessment of knees [31–33].

This research has some limitations. First, we used labelled data with scores assessed by the consensus of two rheumatologists. Validity and generalizability of the model would be improved if a greater number of images with accurate scoring results were offered to the CNN. Second, our model frequently failed to detect intercarpal joints, which are often impaired in RA. For clinical application, these areas need to be included in the assessment by the model. In the present study, it was difficult to identify each area of intercarpal joints with the model, and a model assessing these areas could not be developed. Third, the sensitivity for erosions was obviously low (34.8–42.4%), indicating oversight for erosions by the model. Additionally, PEA for JSN in PIP/IP joints was low (24.3–58.3%) and the specificity was also low (52.0–74.8%), indicating overestimation of JSN by the model. To overcome these problems, a larger quantity of data should be added, and the structure of the network or the parameters of the machine learning need to be considered. We examined several settings of parameters, such as batch size (64, 128, 256 or 512), number of epochs (maximum of 200 epochs), optimization algorithm (Adam, AdaDelta, SGD or RMSprop), and the introduction of batch normalization or dropout. The number of combinations of these settings or parameters is enormous, and further study to optimize them is needed for better performance.

In the present study, we introduced a deep-learning model using CNN, which assesses fine joint destruction of RA. This model provides a partial assessment among many joints that can be destroyed in RA. However, to the best of our knowledge, a CNN-based deep-learning model has not been applied to automated assessment of radiographic joint destruction in RA. The introduction of AI is useful for prevention of oversights, reduction of time and effort, health surveys and assessment by both specialists and non-specialists. In addition, this methodology can be applied to other joints, such as the elbow, shoulder, hip, knee, foot or spine, and to other disorders, such as osteoporosis, fracture or bone tumours. We conclude that image processing with the trained CNN model is promising to assess radiographic finger destruction in RA.

Acknowledgements

We thank Eiryo Kawakami, Haruhiko Koseki, Koiti Asizaki, Takashi Morizono, Tatsuhiko Tsunoda, Ichirou Miki and Toshitada Takemori for planning the research protocol and suggesting constructive opinions, and Eri Itotagawa for collecting the clinical information. T.H. designed the study, constructed the machine leaning model and wrote the initial draft of the manuscript. M.N., N.N. and K.E. contributed to analysis and interpretation of data. J.S., K.S. and A.K. designed the research concept and critically reviewed the manuscript.

Funding: This work was supported by the research program ‘Hub for Predictive and Preventative Precision Medicine Driven by Big Data’ conducted at RIKEN, the Japan Science and Technology Agency (JST) and Osaka University, Council for Science, Technology and Innovation (CSTI), cross-ministerial Strategic Innovation Promotion Program (SIP), and ‘Innovation AI Hospital System’ [Funding Agency: National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN), grant number JPMJIH1504].

Disclosure statement: The authors have declared no conflicts of interest.

Supplementary data

Supplementary data are available at *Rheumatology Advances in Practice* online.

References

1 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.

2 Webb S. Deep learning for biology. Nature 2018;554: 555–7.

3 Chartrand G, Cheng PM, Vorontsov E *et al*. Deep learning: a primer for radiologists. Radiographics 2017; 37:2113–31.

4 Gulshan V, Peng L, Coram M *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.

5 Esteva A, Kuprel B, Novoa RA *et al*. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

6 Fujisawa Y, Otomo Y, Ogata Y *et al*. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019;180:373–81.

7 Coudray N, Ocampo PS, Sakellaropoulos T *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559–67.

8 Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. Lancet Respir Med 2018;6:837–45.

9 Smolen JS, Landewé R, Bijlsma J *et al*. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann Rheum Dis 2017; 76:960–77.

10 Steinbrocker O, Traeger CH, Batterman RC. Therapeutic criteria in rheumatoid arthritis. J Am Med Assoc 1949; 140:659–62.

11 Kellgren JH, Bier F. Radiological signs of rheumatoid arthritis: a study of observer differences in the reading of hand films. Ann Rheum Dis 1956;15:55–60.

12 Sharp JT, Lidsky MD, Collins LC, Moreland J. Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities. Arthritis Rheum 1971;14: 706–20.

13 Sharp JT, Bluhm GB, Brook A *et al*. Reproducibility of multiple-observer scoring of radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis. Arthritis Rheum 1985;28:16–24.

14 Larsen A, Dale K, Eek M. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference films. Acta Radiol Diagn (Stockh) 1977;18: 481–91.

15 Genant HK. Methods of assessing radiographic change in rheumatoid arthritis. Am J Med 1983;75:35–47.

16 Van der Heijde DM, Van Riel PL, Nuver-Zwart IH, Gribnau FW, Van de Putte LB. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;333:1036–8.

17 Arnett FC, Edworthy SM, Bloch DA *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.

18 Aletaha D, Neogi T, Silman AJ *et al*. 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum 2010;62:2569–81.

19 van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26: 743–5.

20 Viola P, Jones M. Rapid Object Detection using a Boosted Cascade of Simple Features. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2001; I-511–8. 2001 Dec 8-14; Hawaii, USA.

21 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv 2014; 1412.6980.

22 Loffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv 2015; 1502.03167.

23 Tokui S, Oono K, Hido S, Clayton J. Chainer: a next-generation open source framework for deep learning. In: Proceedings of Workshop on Machine Learning Systems (LearningSys) in the Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS), Vol. 5; 2015, pp. 1–6.

24 Meyer O, Labarre C, Dougados M *et al*. Anticitrullinated protein/peptide antibody assays in early rheumatoid arthritis for predicting five year radiographic damage. Ann Rheum Dis 2003;62:120–6.

25 Vencovský J, Machácek S, Sedová L *et al*. Autoantibodies can be prognostic markers of an erosive disease in early rheumatoid arthritis. Ann Rheum Dis 2003;62:427–30.

26 Koga T, Okada A, Fukuda T *et al*. Anti-citrullinated peptide antibodies are the strongest predictor of clinically relevant radiographic progression in rheumatoid arthritis patients achieving remission or low disease activity: A post hoc analysis of a nationwide cohort in Japan. PLoS One 2017;12:e0175281.

27 Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One 2017;12:e0178992.

28 Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017; 36:41–51.

29 Larson DB, Chen MC, Lungren MP *et al*. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287:313–22.

30 Lindsey R, Daluiski A, Chopra S *et al*. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci USA 2018;115:11591–6.

31 Liu F, Zhou Z, Samsonov A *et al*. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. Radiology 2018;289:160–9.

32 Tack A, Mukhopadhyay A, Zachow S. Knee menisci segmentation using convolutional neural networks: data from the osteoarthritis initiative. Osteoarthritis Cartilage 2018;26:680–8.

33 Bien N, Rajpurkar P, Ball RL *et al*. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS Med 2018;15:e1002699.