TRANSLATIONAL SCIENCE

Applying cascaded convolutional neural network design further enhances automatic scoring of arthritis disease activity on ultrasound images from rheumatoid arthritis patients

Anders Bossel Holst Christensen ,[1] Søren Andreas Just ,[2] Jakob Kristian Holm Andersen,[1] Thiusius Rajeeth Savarimuthu[1]

Handling editor Josef S Smolen

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/annrheumdis-2019-216636).

[1]Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
[2]Department of Rheumatology, Odense University Hospital, Odense, Denmark

Correspondence to
Anders Bossel Holst Christensen, Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense 5230, Denmark;
abc@mmmi.sdu.dk

Received 12 November 2019
Revised 19 May 2020
Accepted 20 May 2020
Published Online First 5 June 2020

ABSTRACT

Objectives We have previously shown that neural network technology can be used for scoring arthritis disease activity in ultrasound images from rheumatoid arthritis (RA) patients, giving scores according to the EULAR-OMERACT grading system. We have now further developed the architecture of this neural network and can here present a new idea applying cascaded convolutional neural network (CNN) design with even better results. We evaluate the generalisability of this method on unseen data, comparing the CNN with an expert rheumatologist.

Methods The images were graded by an expert rheumatologist according to the EULAR-OMERACT synovitis scoring system. CNNs were systematically trained to find the best configuration. The algorithms were evaluated on a separate test data set and compared with the gradings of an expert rheumatologist on a per-joint basis using a Kappa statistic, and on a per-patient basis using a Wilcoxon signed-rank test.

Results With 1678 images available for training and 322 images for testing the model, it achieved an overall four-class accuracy of 83.9%. On a per-patient level, there was no significant difference between the classifications of the model and of a human expert (p=0.85). Our original CNN had a four-class accuracy of 75.0%.

Conclusions Using a new network architecture we have further enhanced the algorithm and have shown strong agreement with an expert rheumatologist on a per-joint basis and on a per-patient basis. This emphasises the potential of using CNNs with this architecture as a strong assistive tool for the objective assessment of disease activity of RA patients.

Key messages

What is already known about this subject?

► Convolutional neural networks (CNNs) have already been successfully applied to a wide array of health issues.[6] The application of CNNs on ultrasound images for classification of the disease activity of rheumatoid arthritis has enabled four-degree classification with 75.0% prediction accuracy. Previous studies typically used the Kappa statistic as the only measure of agreement between algorithm and human expert.

What does this study add?

► The complexity of automatic grading of rheumatoid arthritis (RA) disease activity into four degrees of severity (using the EULAR-OMERACT classification system) has required a CNN architecture designed specifically for that purpose. In this study we have explored such architectures and designed a new cascade model which has achieved an accuracy of 83.9% on a test set with a naturally occurring distribution of disease activity scores of RA. Furthermore, we have shown that the algorithm performs comparably to an expert rheumatologist on single images and on a collection of images for a single patient.

How might this impact on clinical practice or future developments?

► The new algorithm developed in this study has the potential to provide an operator-independent method for evaluation of disease activity of RA, which can prove beneficial for future trials and clinical practice. Furthermore, the algorithm can potentially be implemented as an assistive tool for the rheumatologist in the clinical practice in the future by analysing multiple images from the same patient, combining data and presenting it to the rheumatologist. This could be both in the context of early disease detection and ensuring sustained remission in patients with established disease.

INTRODUCTION

Systematic power or colour Doppler (CD) ultrasound (US) of joints can be used for early detection of rheumatoid arthritis (RA), predicting radiographic progression and early detection of disease flare in established RA.[1 2] A major problem until recently has been the lack of an internationally recognised system for exactly how to perform RA US scanning and thereafter how to evaluate disease activity on the obtained images. This system has now been developed and named the EULAR-OMERACT Synovitis Scoring (EOSS) system.[1 3 4] The EOSS system uses standardised US scanning positions and describes

To cite: Christensen ABH, Just SA, Andersen JKH, *et al*. *Ann Rheum Dis* 2020;**79**:1189–1193.

Table 1 The colour Doppler ultrasound images in the data set prior to splitting into training, validation and test

Class 0	Class 1	Class 2	Class 3	Total
746	419	305	208	1678

scoring of synovial hypertrophy (SH) from 0 to 3, blood flow in the synovium by Doppler US from 0 to 3 and finally how a combined score from 0 to 3 can be obtained.[1 3 4]

To further mitigate the operator-dependency in scoring disease activity on CD US images in future trials and clinical practice, we proposed the use of convolutional neural networks (CNN) to automatically grade CD US images into four degrees of severity according to the EOSS definitions.[1 5] This study is a continuation of the findings in our previous work, where we managed to develop a CNN for four-class CD US EOSS scoring with a test accuracy of 75.0%.[5]

In recent years, CNN's have been established as the state-of-the-art approach for automatic image recognition and analysis.[6] The capacity of the CNN's to achieve high performance on a wide variety of data originates from their ability to learn the appropriate filters for extracting the information from the data which enables distinguishing between a set of predefined classes (eg, no disease activity/high disease activity) through an iterative optimisation process.[6 7]

In this study, we show that developing a cascade of CNN's, each capable of binary classification, resulting in an algorithm capable of four-class classification, is a viable method for automated grading of ultrasound images from RA patients, performing comparably to a human expert.

METHODS

Materials

The data used in this study is the same as that which was used for our previous study, that is, 1678 CD US images, which all came from an RA study (ClinicalTrials.gov: NCT0262299). Here 40 patients with RA (20 patients with long-standing disease >5 years and 20 patients with early untreated disease) were followed for 6 months, as previously described.[8] During this period, synovial biopsies were performed from the wrist at baseline and 6 months, and US scans of the hand where synovial biopsies were taken from, were performed at baseline, 3 months and 6 months. For patient baseline characteristics and treatment, see online supplementary material, table D1. For detailed data on MRI, US and synovial biopsy scoring and changes during the trial, please see Just *et al* 2019.[8]

As according to the EOSS guidelines, the joints that were scanned included the radiocarpal-intercarpal joint, the radioulnar joint, the proximal interphalangeal (PIP) joints and the metacarpophalangeal (MCP) joints from the dorsal side of the hand.

In addition to the 1678 images captured during this study (all CD US images, from all visits), 322 CD US images were captured by the same rheumatologist (SAJ) from 14 other patients not in the described trial, to be used for testing the algorithm. All patients had given their consents prior to the acquisition of the images.

All images were captured using a General Electric Logiq 9 US machine and a linear array ML6-15 transducer. US machine settings were unchanged throughout the study, with CD signal gain set to a sensitivity just below the disappearance of colour noise. All images were anonymised. A rheumatologist (SAJ) with approximately 9 years of experience with US scanning scored US images in accordance with the EOSS system. Scoring SH from 0 to 3, blood flow in the synovium by CD from 0 to 3, thereby giving the combined EOSS score from 0 to 3.[1] Examples of the CD US images can be found in online supplementary material E as heatmaps.

Patients and public involvement

Patients or the public were not involved in the design, conduct, reporting or dissemination plans of our research.

Procedure

In this study we trained CNN's for two separate cases. One in which data augmentation was used (case 2) and one in which it was not (case 1).

For both cases, we randomly split the original data set (table 1) into a training set containing 80% of the data and validation and test sets each containing 10% of the data.

In case two, after splitting the data into training, validation and test sets, the distribution of the classes in the training set was balanced, and data augmentation techniques were applied, resulting in training sets four times their original sizes. Data augmentation is a regularisation technique used in deep learning for augmenting the data set with predictable transformations of the data, which often improves the model's generalisability to

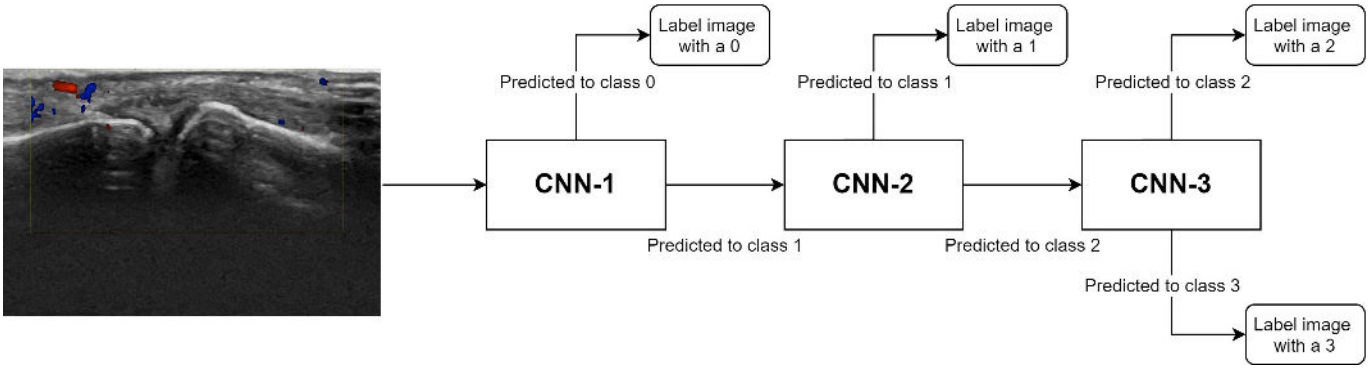

Figure 1 Illustration of how the cascade model works. CNN-1 is the first neural network in the model and is capable of distinguishing between class 0 and classes 1, 2 and 3. CNN-2 is the second neural network in the model and is capable of distinguishing between class 1 and classes 2 and 3. Finally, CNN-3 is capable of distinguishing between classes 2 and 3. Each CNN outputs probability scores for each of their classes. The prediction of a CNN is the class associated with the highest probability score. If CNN-1 predicts the input image to be of class 0, that classification will be accepted. If, on the other hand, CNN-1 predicts it to be of class 1, 2 or 3, CNN-2 will attempt to classify the image to class 1 or classes 2 or 3, in which case CNN-3 will make the final prediction. CNN, convolutional neural network.

Table 2 The colour Doppler ultrasound images in the test set for testing the cascade model

Class 0	Class 1	Class 2	Class 3	Total
225	53	24	20	322

unseen data. See online supplementary material B for figures explaining the process.

In each of the two cases, six CNN's each capable of binary classification were built and trained; three built and trained from scratch, and three built and trained on features extracted from Inception-v3's 'mixed-4'-layer pretrained on the ImageNet data (see online supplementary material F for network architectures). Of these 12 CNN's, three were needed to enable the cascade model to distinguish between all four (0 to 3) EOSS scores (ie, classes), see figure 1. Training, validation and test sets were created separately for each binary CNN. These data sets only contained images of the classes they should learn to classify (eg, the last CNN in the cascade model was only trained on class 2 and class 3 images). After every epoch of training, for a total of 120 epochs, performance was measured on the validation set, which was also used to optimise the hyperparameters controlling how the CNN's learn. Because this optimisation process may cause the CNN's to overfit to the validation sets,[9] the test sets were used at the end to test the performance of the CNN's on entirely new data.

Prior to beginning the training process, the data was preprocessed, which involved cropping and resizing the images to size 299×299, as well as performing zero-centering and normalisation of the images.

The training process was performed using the high-level application programming interface Keras with the TensorFlow backend for numerical computations.[10 11]

The performances of the best performing CNN for each of the 12 CNN's are shown in online supplementary material C.

Cascade model

The working principle of the cascade model is presented in figure 1. The selection of the three best-performing CNN's involved comparisons of the accuracies on the validation sets and on a per-class basis as well as comparisons of their performance in a cross-validation test.

Statistics

For testing the performance of the cascade model, the additional 322 CD US images from 14 RA patients were generated and graded (table 2). These images served as the test set and were used to determine the overall accuracy as well as the per-class accuracy of the cascade model. Typically, in machine learning tasks, the distribution of the classes in the test set is balanced. For this study however, we wanted to test how the cascade model would perform if it were to be implemented in a typical hospital setting. For that reason, the distribution of the data in the test set resembles the expected distribution of disease severity of RA among the patients at section of Rheumatology, Svendborg Hospital - Odense University Hospital.

Cohen's Kappa statistic was used on the test set to evaluate the agreement between an expert rheumatologist and the cascade model on a per-joint basis. Furthermore, sensitivity and specificity were calculated for each binary classification network in the cascade, with the rheumatologist (SAJ) as golden standard.

For evaluating the agreement between the expert rheumatologist (SAJ) and the cascade model on a per-patient basis, we combined the EOSS scores of 24 joints (MCP1-5, PIP1-5, radioulnar and radiocarpal-intercarpal joints for both hands) into a composite score calculated by the sum of the EOSS scores of the individual joints, resulting in a score ranging from 0 to 72 for a single patient. A Wilcoxon signed-rank test was performed to test for any significant difference between the composite scores of the expert rheumatologist (SAJ) and of the cascade model. Wilcoxon was used as we compare two repeated measurements on a single sample (all CD US from both hands), and the distribution of the differences between the two measurements cannot be assumed to be normally distributed. For data on interclass correlation and validity of each separate CNN, see online supplementary material A and C.

RESULTS

Convolutional neural networks for binary classification

The top three performing CNN's selected for the cascade model were all built as simple, dense top classifiers trained on features extracted from the 'mixed-4'-layer of Inception-v3 pretrained on the ImageNet data (see the figures in online supplementary material F). CNN-1 and CNN-3 were trained in case 1, whereas CNN-2 benefitted from the data augmentation in case 2. As summarised in table 3, the individual CNN's in the cascade achieve close to 90% accuracy on the test set when distinguishing between EOSS score 0 and EOSS scores 1, 2 and 3 (CNN-1), and when distinguishing between EOSS score 1 and EOSS scores 2 and 3 (CNN-2). CNN-3 is making several false negatives (predicting EOSS 3 as EOSS 2), reaching just 78.9% accuracy on the test set. Confusion matrices underlying the numbers shown in table 3 can be found in online supplementary material A.

The cascade model

Testing the cascade model on the 322 CD US images, the model achieved a four-class accuracy of 83.9%. As shown in table 4, the cascade model performs significantly better on images with an EOSS score of 0 and 3 compared with images with an EOSS score of 1 and 2. Considering images of EOSS score 0 and 1 negatives and images of EOSS score 2 and 3 positives, the sensitivity and specificity of the cascade model are 95.5% and 97.5% respectively. This emphasises the fact that the majority of the

Table 3 Test results of the three CNN's for binary classification included in the cascade model

Network	Test accuracy	Sensitivity	Specificity	AUC
CNN-1	89.9%	90.5%	89.3%	0.96
CNN-2	88.3%	91.5%	85.1%	0.94
CNN-3	78.9%	57.7%	100.0%	0.93

AUC, area under the (receiver operating characteristic) curve; CNN, convolutional neural network.

Table 4 Predictions of the cascade model versus gradings of the rheumatologist

	Cascade model					
Rheumatologist	0	1	2	3	Total	Accuracy(%)
0	211	12	2	0	225	93.8
1	20	28	5	0	53	52.8
2	0	2	15	7	24	62.5
3	0	0	4	16	20	80.0
Total	231	42	26	23	322	

misclassifications of the cascade model are between EOSS scores 0 and 1 and between EOSS scores 2 and 3. Using Cohen's Kappa statistic as a measure of agreement between the grades of the rheumatologist and the predictions of the cascade model, the unweighted Kappa score was 0.65, and the linearly weighted Kappa score was 0.79, both indicating good agreement on a per-joint basis.[12] For comparison, the interobserver agreement among 12 US-experienced rheumatologists was found by Terslev *et al*[3] to be excellent, with a Kappa score of 0.86.

Using a Wilcoxon signed-rank test on the per-patient composite EOSS scores estimated by the rheumatologist and by the cascade model, with an alpha level of 0.05, no significant difference (p=0.85) between the composite EOSS scores of the rheumatologist and of the cascade model was found.

DISCUSSION

We herein presented some of the work we have done since our last contribution in improving CNN technology for automatic classification of disease activity on ultrasound images from RA patients according to the EOSS system.[5] With 2000 CD US images available for training and testing, we have shown that dividing a four-degree classification task into three successive binary classification tasks has resulted in a model capable of making correct classifications in 83.9% of the cases for a test set of ultrasound images with a naturally occurring distribution of RA joint disease activity scores. Unfortunately, due to differences in the way the data was split for training, validation and test in this study and our previous study, the cascade model was not tested using the same test set as the network developed in our previous study.[5]

Second, we have shown that with a relatively small data set, on this specific classification task, using feature extraction from Inception-v3 is a better strategy than building convolutional neural networks and training them from scratch. Third, we have shown that the majority of misclassifications made by the cascade model happen between images with EOSS scores 0 and 1 and between images with EOSS scores 2 and 3. Finally, we have found that any such misclassifications seem to get evened out when predicting composite EOSS scores on a per-patient basis, emphasising the applicability of the cascade model for predicting disease activity based on a set of multiple joint pairs.

We expect that the use of CNN's for automated classification across different health facilities and countries could provide more comparable, unbiassed gradings for use in future studies. Furthermore, with future projections of an ageing population and an increasing demand for rheumatologists, the use of CNN's as an assistive tool for rheumatologists could help to meet the increasing demand and potentially assist in either finding signs of early disease or finding signs of disease flare in established disease.[13–16]

Improvements in the performance of the classification algorithms are continuously being made. Potential areas for future work include training CNN's to make less misclassifications on images with an EOSS score of 1. Also, for further improving the generalisability of the algorithm, model ensembling can be explored, as many studies have had success with averaging predictions across several different types of models.[17 18]

While the cascade model presented in this study was trained to classify CD US images, a similar algorithm for classification of greyscale US images is being developed.

One of the main limiting factors for achieving better performance with the CNN's is the amount of data available to use for training. However, with increased awareness about this limitation among the developers and among the healthcare professionals, we expect more data to be generated in the future, and a corresponding increase in the performance of the CNN's.

A potential limitation of this study is that CD US images were graded by a single rheumatologist. In the optimal case, one would have multiple highly experienced rheumatologists grading the images and use their scores by either (1) assigning the image with the grade that is most frequent among the experts, potentially 'evening out' the effect of mistakes or (2) excluding the images with low agreement among the graders, for example, with less than four out of five graders agreeing on the label for the image. The downsides of (2) would be that data is already scarce and keeping only the images that are 'easiest' to grade could risk making the network unable to classify the images that are 'harder' to classify.

We did not find that erosion or osteophytes affected scoring (online supplementary material E), but it is a limitation that only few patients in the study had severe mutilating RA or osteoarthritis (OA) joint destruction. We are therefore now developing an OA neural network (NN) and collecting US images of joints with severe RA destruction for NN training.

A further limitation is also; we have not systematically tested how robust the cascade model is to CD US images taken slightly out of plane from optimal scan position. The transformations that are made to the images in this study, through data augmentation techniques, have further increased the generalisability of the algorithm and its viewpoint-invariance and scale-invariance. We therefore do not believe it is a problem, although we will test it in future studies. Further, the impact of using different US systems needs to be tested, as images in this study were all obtained from the same system.

In conclusion, this study further emphasises the potential of using CNN technology for automated classification of disease activity on US images of RA patients using the EOSS system. By optimising the CNN design, we have developed a model that achieves an accuracy of 83.9%. We have further shown that when combining the data from several joints of the same patient, the algorithm does not score significantly different than an experienced rheumatologist. The current developed CNN needs more training data from different centres scored by different experts to eliminate any potential bias in the training data, to ensure the effect of viewpoint-setting and scale-setting invariance and any potential effects of OA and erosions on scoring. We are currently working on all these developing points. We still believe this technology, especially combining US data from several joints of the same patient, analysing them and presenting the data to the clinician in a fast and unbiased manner, could prove a valuable assistive tool for the assessment of disease activity of RA patients in both daily clinical practice and in future trials.

Twitter Søren Andreas Just @JustSoren

Acknowledgements Our gratitude to all the study subjects that have agreed to let us capture and use their ultrasound images for training of these algorithms.

Contributors Development, training and evaluation of algorithms: ABHC. Collection and grading of data: SAJ. Design of the study: ABHC, TRS, SAJ. Counselling: TRS, SAJ, JKHA. Drafting, revising and final approval: ABHC, SAJ, TRS.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Disclaimer This study was done as a master thesis project at the Maersk Mc-Kinney Moller Institute, University of Southern Denmark, and has not received any funding.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval The SynRA study, where ultrasound images are from, is approved by the regional ethics review board (S-20140062) and the Danish data protection agency (2008-58-0035). All participants gave oral and written consent to participate.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iDs
Anders Bossel Holst Christensen http://orcid.org/0000-0002-6053-0287
Søren Andreas Just http://orcid.org/0000-0002-3946-5919

REFERENCES

1 D'Agostino M-A, Terslev L, Aegerter P, *et al*. Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce-Part 1: definition and development of a standardised, consensus-based scoring system. *RMD Open* 2017;3:e000428.
2 Paulshus Sundlisæter N, Aga A-B, Olsen IC, *et al*. Clinical and ultrasound remission after 6 months of treat-to-target therapy in early rheumatoid arthritis: associations to future good radiographic and physical outcomes. *Ann Rheum Dis* 2018;77:1425.
3 Terslev L, Naredo E, Aegerter P, *et al*. Scoring ultrasound synovitis in rheumatoid arthritis: a EULAR-OMERACT ultrasound taskforce-Part 2: reliability and application to multiple joints of a standardised consensus-based scoring system. *RMD Open* 2017;3:e000427.
4 Hammer HB, Bolton-King P, Bakkeheim V, *et al*. Examination of intra and interrater reliability with a new ultrasonographic reference atlas for scoring of synovitis in patients with rheumatoid arthritis. *Ann Rheum Dis* 2011;70:1995–8.
5 Andersen JKH, Pedersen JS, Laursen MS, *et al*. Neural networks for automatic scoring of arthritis disease activity on ultrasound images. *RMD Open* 2019;5:e000891.
6 Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
7 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44.
8 Just SA, Nielsen C, Werlinrud JC, *et al*. Six-month prospective trial in early and long-standing rheumatoid arthritis: evaluating disease activity in the wrist through sequential synovial histopathological analysis, RAMRIS magnetic resonance score and EULAR-OMERACT ultrasound score. *RMD Open* 2019;5:e000951.
9 Kuhn M, Johnson K. *Applied predictive modeling*. New York: Springer, 2013: 67.
10 Keras CF, 2015. Available: https://keras.io
11 OSDI. Tensorflow: a system for large-scale machine learning; 2016.
12 Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;33:363–74.
13 Combe B. Progression in early rheumatoid arthritis. *Best Pract Res Clin Rheumatol* 2009;23:59–69.
14 Zink A, Braun J, Gromnica-Ihle E, *et al*. Memorandum der deutschen gesellschaft für rheumatologie zur versorgungsqualität in der rheumatologie—update 2016. *Zeitschrift für Rheumatologie* 2017;76:195–207.
15 Deal CL, Hooker R, Harrington T, *et al*. The United States rheumatology workforce: supply and demand, 2005-2025. *Arthritis Rheum* 2007;56:722–9.
16 Battafarano DF, Ditmyer M, Bolster MB, *et al*. 2015 American College of rheumatology workforce study: supply and demand projections of adult rheumatology workforce, 2015-2030. *Arthritis Care Res* 2018;70:617–26.
17 De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
18 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *NIPS*, 2012.