nature reviews rheumatology

https://doi.org/10.1038/s41584-023-01074-5

Review article

Check for updates

Deep learning in rheumatological image interpretation

Berend C. Stoel[1], **Marius Staring**[1], **Monique Reijnierse**[2] & **Annette H. M. van der Helm-van Mil**[3]

Abstract

Artificial intelligence techniques, specifically deep learning, have already affected daily life in a wide range of areas. Likewise, initial applications have been explored in rheumatology. Deep learning might not easily surpass the accuracy of classic techniques when performing classification or regression on low-dimensional numerical data. With images as input, however, deep learning has become so successful that it has already outperformed the majority of conventional image-processing techniques developed during the past 50 years. As with any new imaging technology, rheumatologists and radiologists need to consider adapting their arsenal of diagnostic, prognostic and monitoring tools, and even their clinical role and collaborations. This adaptation requires a basic understanding of the technical background of deep learning, to efficiently utilize its benefits but also to recognize its drawbacks and pitfalls, as blindly relying on deep learning might be at odds with its capabilities. To facilitate such an understanding, it is necessary to provide an overview of deep-learning techniques for automatic image analysis in detecting, quantifying, predicting and monitoring rheumatic diseases, and of currently published deep-learning applications in radiological imaging for rheumatology, with critical assessment of possible limitations, errors and confounders, and conceivable consequences for rheumatologists and radiologists in clinical practice.

Sections

[1]Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands. [2]Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands. [3]Department of Rheumatology, Leiden University Medical Center, Leiden, the Netherlands. e-mail: b.c.stoel@lumc.nl

Key points

- The number of research studies on deep learning in rheumatological imaging has grown rapidly during the past 5 years, but they mainly consist of pilot studies that require external validation.
- Confounding factors and errors in deep-learning methods need to be ruled out before deep learning can be applied in clinical practice, for which the intended use should be strictly defined.
- Deep-learning techniques, together with mapping to explain their reasoning, will enable hypothesis-free image analysis and could identify new imaging biomarkers.
- Deep learning might assist rheumatologists and radiologists in interpreting rheumatological images, increasing their diagnostic, prognostic and monitoring accuracy, and decreasing workloads and costs.

Introduction

Although the term 'artificial intelligence' (AI) is used as a synonym for techniques that were developed only very recently, it actually has a long history, generally running parallel with the development of computers. However, with advances in AI techniques that use artificial neural networks, it has now become possible to train complex networks, which is referred to as 'deep learning', on enormous datasets. Deep-learning technology has now gained such momentum, with such a variety of successful applications and implications in daily life, that even AI experts are sometimes surprised.

Applications of deep learning are starting to be explored in rheumatology[1]. For analysis of low-dimensional clinical data, such as a series of single blood biomarkers, or demographic data, deep learning might not easily surpass the accuracy of statistical models[2]. By contrast, with images as input, deep learning has already outperformed most conventional AI techniques that are mainly based on manually designed computer programmes performing specific tasks to detect anatomical structures and quantify predefined features from these objects: the so-called imaging biomarkers. These methods have culminated in an approach, ambitiously called radiomics, in which a large number of predefined imaging biomarkers are calculated from an image to identify a subset of features that contain the correct information for a specific task. In the past 5–10 years, however, this paradigm has shifted dramatically from human-designed computer programmes towards the extraction of machine-learned features. Why this shift has become so instrumental in accurately interpreting images is illustrated in Fig. 1, which is based on a well-known optical illusion that demonstrates the strengths and weaknesses of both human and computerized visual interpretation.

The use of deep learning in the various tasks involved in image analysis has made remarkable progress, possibly outperforming human image interpretation. However, to prevent prematurely embracing these new AI technologies or being too reluctant in making use of their added value, a basic understanding of deep learning and its possible confounders and pitfalls is needed, to recognize and acknowledge both the benefits of deep learning and its (current) limitations. In this Review we therefore provide an overview of emerging deep-learning applications in the interpretation of rheumatological images. We focus on post-processing in specifically radiological imaging, although deep learning is also being applied in image reconstruction[3], more general musculoskeletal imaging and in post-processing in non-radiological imaging (such as histology). First, we give an overview of the different AI tasks in image processing and explain the basic principles of convolutional neural networks (CNNs). Then, we provide an overview of the current literature on deep learning in rheumatological image interpretation, followed by a discussion of possible confounders and pitfalls in these studies, and finally an outlook on the possible influence of deep learning in clinical practice for rheumatologists and (musculoskeletal) radiologists.

AI tasks in image analysis

In this Review, we use the framework presented in Fig. 2 to categorize the different AI tasks involved in automatically extracting information from medical images. Such a task could be to automatically delineate specific objects or specify regions in the image (defined here as 'detection'), optionally followed by quantification of an imaging biomarker; to assign a class label or category to the entire image (image 'classification'); or to estimate an ordinal or scalar number from a region or entire image ('regression'). Within each of these three main tasks, there can be different targets.

For a detection task, targets can be regions of interest (ROIs), such as a rectangular area in an image, a slice in a 3D volume, or a frame within a sequence of images (for example, to automatically select the most appropriate frame in ultrasonographic images)[4]; anatomical structures (for example, detecting tibial cartilage from MRI) prior to quantification[5]; or lesions or pathological regions (for example, detecting osteophytes from radiographs of the hand)[6]. The detection of all pixels that belong to a particular anatomy or lesion is also known as (semantic) segmentation. After segmentation, features can be measured (by quantification) to derive a particular category of patient outcome (Fig. 2, route A).

Image classification can have targets to distinguish between categories such as patient groups[7] (versus healthy individuals), between treatment arms[8], or between rapid and slow decliners, responders and non-responders and other comparisons. This approach does not require any explicit segmentation or quantification steps: the input is an image and the output is a class label (Fig. 2, route B).

Finally, a regression task can be to estimate or simulate visual scoring (for example, to automatically perform Kellgren–Lawrence scoring on knee radiographs)[9]; to calculate the risk of an event over time, such as the need for a future total knee replacement (prediction)[10]; or to estimate other non-imaging biomarkers (for example, to estimate functional outcomes from images, to study function–structure relationships[11], or to localize symptoms such as pain)[12]. After finishing this regression task, the patient outcome category can be determined by a subsequent classification step (Fig. 2, route C).

In the above examples, a single image (or a set of images from different image modalities) is taken as input for a deep-learning model (for cross-sectional analyses). However, the same framework can be used for longitudinal analyses, where time series of images are taken as input for the model to detect changes over time, and/or to perform classification or regression on these changes.

Basic principles of convolutional neural networks

There are many online tutorials available that explain the foundations of deep learning, among which the CS230 Deep Learning tutorial is recommended, and a glossary of common machine-learning terminology can also be found online. Here, we discuss the most relevant topics.

Layers of neurons

Essentially, an artificial neural network consists of units, called neurons, that are organized in layers, with connections between them from preceding to following layers. A neuron becomes activated when the total of input signals (representing image intensities or activations from connected neurons) reaches a threshold. Activations propagate through the entire network until activation in the final output layer can be interpreted, for example, as a disease label, anatomical label or disease activity score. The depth of a neural network refers to the number of layers it contains, hence the name 'deep learning'. A key difference from classic machine-learning techniques is that in neural networks the general sensitivity of a neuron (called bias) and its sensitivity to individual input signals (defined by feature weights) can be tuned automatically, whereas in classic machine learning these features are pre-defined. During this tuning (training), groups of neurons form features that are useful for reaching a training objective. So-called 'CNNs' are specialized in the extraction of features from images.

Extracting local features from images: the concept of convolution

The core function of deep learning in image processing is to extract relevant information from any location within the image. The way in which local information is extracted should be independent of where this information can be found in the image; for example, a knee with signs of osteoarthritis (OA) should be detected irrespective of where it is placed in the MRI scanner's field of view. Commonly, this concept has been implemented in deep learning by so-called convolutional filters (Fig. 3). A square kernel, divided into cells that are assigned weights, 'hovers' over the image and at each location the weights in the kernel are multiplied by the corresponding intensities in the image. The weighted sum is then recorded in the output image. Selection of the weights defines the type of local feature that is quantified, such as the average intensity (Fig. 3b), horizontal transitions from dark to bright areas (Fig. 3c) or dark objects surrounded by a brighter background (Fig. 3d).

In the illustration shown in Fig. 3, the kernel weights were preselected, whereas in CNNs they are 'learnable', which means that the weights in the kernels undergo optimization to perform a specific task, such as the detection and/or quantification of inflamed tissue at any location in the image. Generally, many (combinations of) different kernels are needed to learn these complex tasks, either trained from scratch or initialized by pre-training for general detection tasks, such as interpreting photographs from daily life. After training, the first layers of kernels usually turn out to measure basic features as illustrated in Fig. 3, and in subsequent layers kernels seem to quantify higher-order composite features that are needed to locate these inflammatory patterns.

Coaching and training schedules

The learning process is often categorized into supervised, unsupervised and reinforcement learning. Supervised learning is arguably the most popular paradigm, and it requires an often manually defined target (the human supervision), such as manual annotations, or clinical (risk) scores for each image in the training set. Unsupervised learning does not require these labels, and is aimed at discovering patterns in the data without explicit guidance, which often proves less powerful than supervised learning. In reinforcement learning, a model learns through trial and error, optimizing its features based on rewards and penalties. The latter has been successful in game playing, but is less common in the medical imaging domain. Most studies considered in this Review are based on supervised learning.

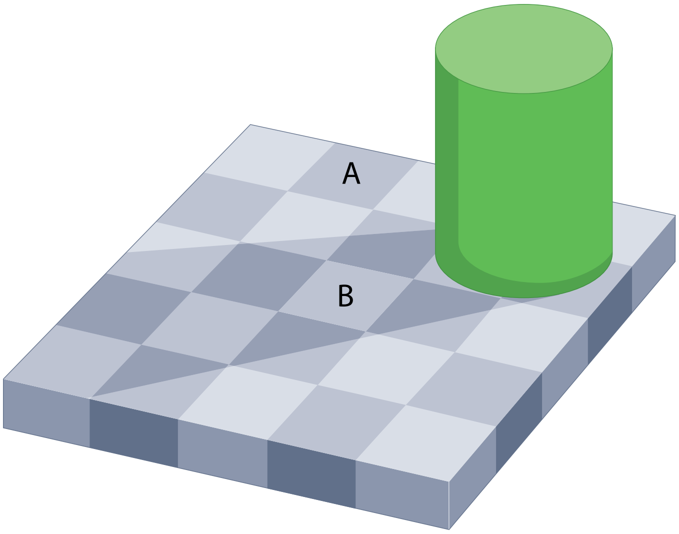

Fig. 1 | Optical illusion illustrating the strength and weakness of both computer and human visual systems. The fact that squares A and B are perceived as dark and bright areas, respectively, illustrates the ease with which a human inevitably interprets the whole scene, instantly recognizing a cylinder casting a shadow over the checkerboard. The figure also reveals the human inability to perceive absolute image intensities, as squares A and B have exactly the same grey value (this can be checked by masking the remaining part of the image). Humans are, therefore, intrinsically limited in truly quantifying images. In the past 50 years, conventional artificial intelligence techniques have failed to automatically detect these types of objects, because pixels were usually classified based on predefined local features, which will put squares A and B into the same category. Proper interpretation requires higher-level representations of this checkerboard, the cylinder, its shadow and even the light source, which is not present in this picture. With the advent of deep-learning techniques this high-level interpretation has become available, at the same time making use of the fact that computers can quantify absolute intensities more accurately than humans. Reprinted from http://persci.mit.edu/gallery/checkershadow, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Training a CNN is nothing more than mathematically optimizing the weights and biases of all neurons, such that the distance to the training objective (the so-called loss function) is minimized. This training is an iterative process, with each iteration making small improvements to the minimization. In every iteration only a randomly selected part of the training data (a batch) is shown to the network, which helps to keep memory requirements in check and possibly to accelerate the training process. After a number of iterations, the network will have seen all the data exactly once, which is referred to as an epoch. Typically, multiple epochs are needed before reaching an optimal outcome where new iterations no longer yield an improvement in performance. Depending on the task at hand, a specific loss function is chosen, which greatly influences the final performance. For segmentation, a combination of cross-entropy and the Dice coefficient is a popular choice, whereas for classification tasks cross-entropy is typically used, and for regression tasks the mean-squared error might be more suitable. This choice needs to be evaluated during the design stage.

CNNs also possess so-called hyper-parameters that are fixed during training, but still need to be optimized by the designer of the network. Examples include the number of layers, kernel sizes, learning rates, but also the specific choice of loss function or optimization routine.

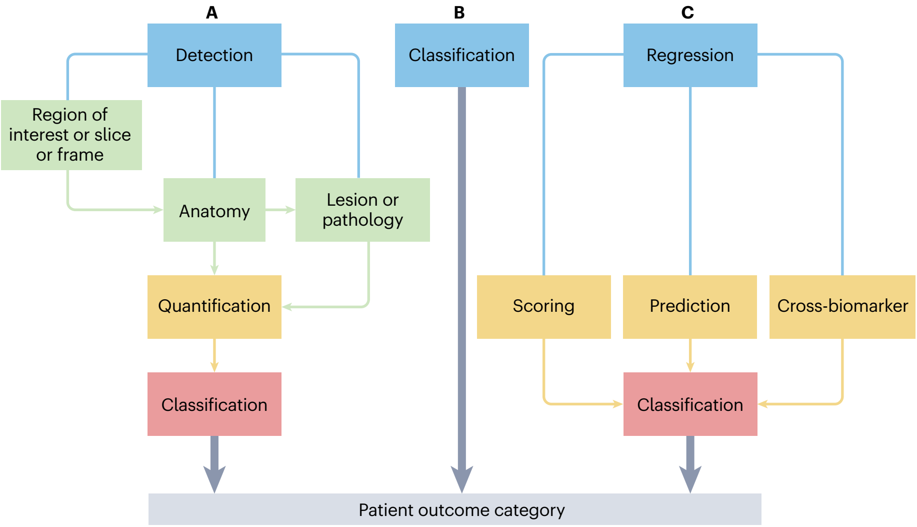

Fig. 2 | Different routes to calculating an outcome measure or category from an image. In route A, regions, anatomy or lesions are first detected, optionally followed by quantification of these objects, based on which a classification can be performed to represent a patient outcome category. In route B, classification into a patient outcome is performed directly based on the image. In route C, images are first quantified by deep learning to produce an automatic score, a prediction or risk assessment of a certain event, or a different (non-imaging) biomarker, optionally followed by a classification.

Training, validation, testing and external datasets

For suitable training and validation of a model, datasets need to be carefully defined and selected. A training set should be used for network optimization, as described above. A separate validation dataset enables the evaluation of network performance and overfitting, and the selection of hyper-parameters. A test set, distinct from but similar to the training and validation sets, enables determination of the final model's performance. An external test set, coming from a different source (for example, another patient population, imaging model or manufacturer, imaging protocol, or hospital), but with the same intended use, serves to guarantee the generalizability of the developed method.

Explainability: insight into the black box

Although deep-learning methods have proven very effective, because they are essentially 'black boxes', they do not explain themselves, and only give limited insight into the factors that influence their decisions. In detection or segmentation tasks, a network's result might be sufficiently self-explanatory, but for classification and regression tasks this lack of explanation can become problematic. For example, when automatically grading the severity of knee OA following the Kellgren–Lawrence scoring system, it might be important that not only a calculated grade is returned by the system but also the radiographic features that are at the basis of this automatic grading, in conjunction with an explanation that is relatable to the physician or even the patient. Such an output would increase trust in these systems, improving justification for their use by physicians, who are ultimately held accountable for clinical decision-making. Although methods to improve explainability are still being developed, 'saliency maps' are a popular way of visually pinpointing image regions that are important for a network's decision-making[13]. They are often displayed as heat maps, and present a visual explanation of contributions to classification, facilitating scrutinization of the network's output.

Overview of the current literature on image interpretation

As this is a multidisciplinary topic with literature from various sources and with rapidly evolving AI terminology, we iteratively used broad search terms such as "image processing", "artificial intelligence" and "rheumatology", and subsequently excluded manually all AI methods that were not using artificial neural networks or that were applied in general musculoskeletal radiology. For medical sources we mainly consulted PubMed, Web of Science, Google Scholar and Scopus, and for technical papers CiteSeerX, Semantic Scholar, ArXiv and GitHub.

The number of publications on deep learning in the field of image interpretation in rheumatology seems to be growing exponentially, with only a single study per year in 2011 on preliminary results[6], increasing to two publications per month in 2023 (Fig. 4a). Among 80 studies of deep learning published up to June 2023, OA, rheumatoid arthritis (RA) and spondyloarthritis (SpA) were the most frequently studied rheumatic diseases (Fig. 4b). By imaging modality, most studies in OA used MRI or plain radiography, and studies in RA predominantly used plain radiography, ultrasonography and MRI, reflecting clinical practice and research methodology, as AI researchers generally use existing data from clinical trials, for which the imaging protocols have already been defined. The studies can also be classified according to their use of the various tasks of detection, classification and regression (Fig. 4c), and according to the anatomical sites that were analysed (Fig. 4d and e).

Deep learning in OA

Among the rheumatic diseases, OA was traditionally the most active research field for the use of conventional AI. As a consequence, deep learning in OA is currently also the most active field for the deep-learning tasks of detection, classification and regression. Predominantly driven by the OA Initiative[14], which has provided open access to training data from MRI and plain radiography of the knee, many deep-learning methods have been applied to detection, with some studies also developing classification and regression methodology.

In one approach using plain radiographs of the knee, quantification of OA severity was accomplished by first detecting the ROI containing the knee joint, and then applying a regression neural network to perform automatic Kellgren–Lawrence grading in that specific region[9,15–19]. Deep learning has also been applied to prediction of whether a patient will need total knee arthroplasty within 9 years or 5 years, using plain radiographs alone[10] or in combination with MRI and clinical data[20]. This use of information from different sources is referred to as multimodal deep learning. Other researchers have used multimodal methods to predict progression of knee OA, based on plain radiography and clinical data, with auxiliary training of Kellgren–Lawrence grading[21]. A competition in a so-called 'grand challenge' has been organized to encourage scientists and engineers to develop AI methods for the prediction of symptomatic radiographic knee OA in 78 months, based on plain radiography, MRI and clinical data[22]. Treatment effects of knee joint distraction have also been evaluated on the basis of commercially available deep-learning software for analysis of knee radiography[23].

Employing MRI scans, OA quantification has mainly followed the detection route (Fig. 2, route A), by first segmenting bones and cartilage and then quantifying these segments to estimate OA severity. Initial deep-learning research was done on solely the detection of cartilage and/or (cortical) bone[5,24–27], as this segmentation is a challenging task in itself. Subsequently, these MRI segmentation results were available to enable the use of morphology (such as cartilage thickness or bone flattening) or relaxometry (T2 mapping) to quantify knee OA severity[28–31]. Segmentations are also used as a basis for classification, for example, the presence or absence of meniscal or patellar cartilage lesions[32], or for quantification of femoral and tibial cartilage degeneration[33]. Grading of lesions based on knee MRI has been explored to (for example) automatically assess cartilage and bone marrow lesions[34] or to stratify patients into morphological phenotypes[35].

An interesting development is the use of a neural network to predict non-imaging biomarkers or patient-reported outcome measures, based on imaging only (a task we called 'cross-biomarker regression'). For example, knee pain can be estimated from MRI data by a regression network[12]. In itself, the clinical relevance of such a technique may be somewhat limited. Once the resulting saliency maps can be interpreted reliably, however, specific sources of knee pain could be studied.

Apart from knee OA, deep-learning methods also found their way into the detection of wrist cartilage from MRI data[36], the detection of osteophytes in hand radiographs[6], the identification of hip OA[37,38], and automatic hip OA grading based on plain radiographs[39].

In addition to the above post-processing tasks, deep learning can support radiologists and technicians during image acquisition procedures. In ultrasonography of the hand, for example, deep learning can be used to automatically select the most informative frame for assessing the metacarpal head[4].

Deep learning in RA

In RA, most of the deep-learning research has been focussed on analysing plain radiographs, ultrasonography data and MRI data relating to the hands, wrists and feet. Plain radiography analysis encompasses mainly the detection or grading of bone erosions in the phalanges,

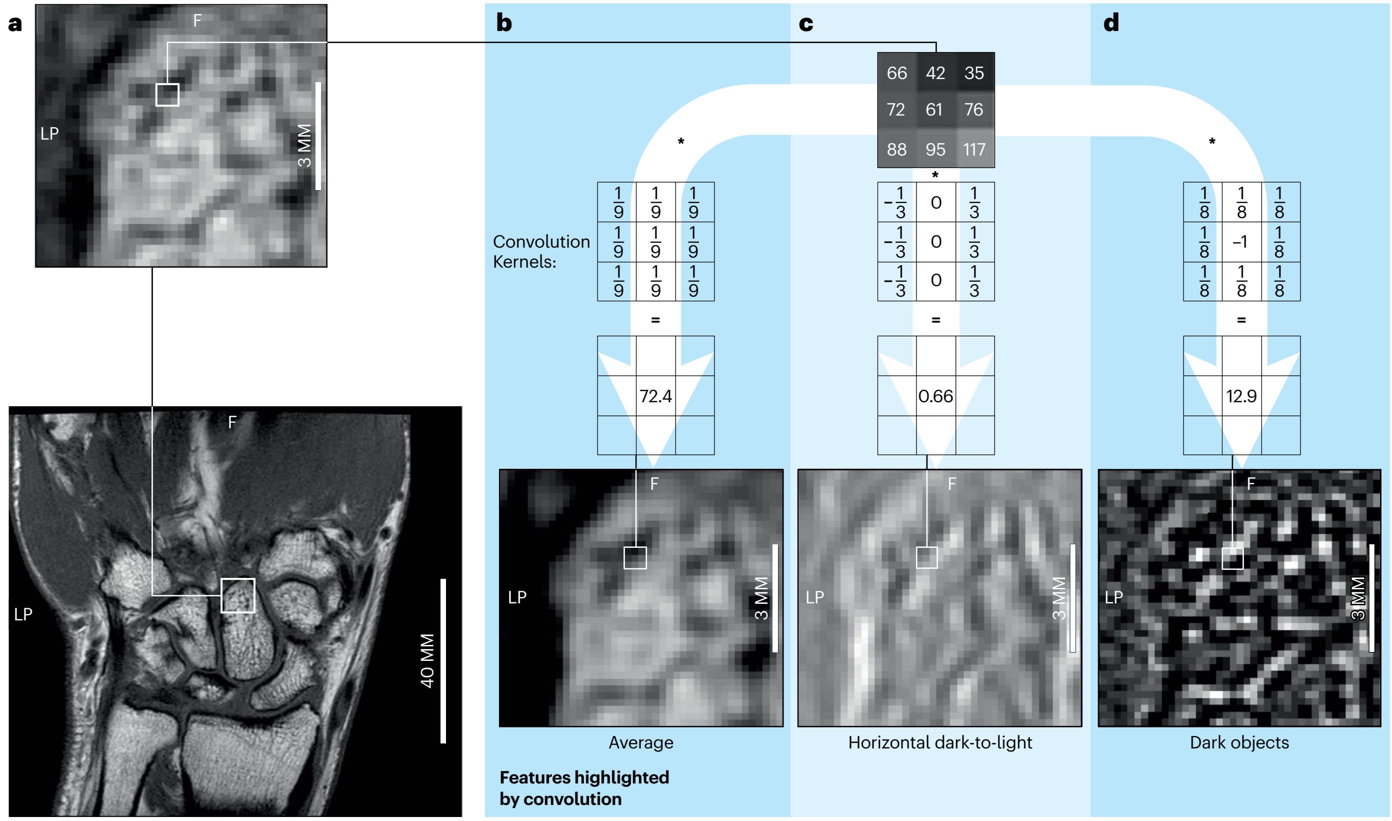
a
F
LP
3 MM
F
LP
40 MM
b
c
d
66 42 35
72 61 76
88 95 117
*
*
*
Convolution Kernels:
$\frac{1}{9}$ $\frac{1}{9}$ $\frac{1}{9}$
$\frac{1}{9}$ $\frac{1}{9}$ $\frac{1}{9}$
$\frac{1}{9}$ $\frac{1}{9}$ $\frac{1}{9}$
$-\frac{1}{3}$ 0 $\frac{1}{3}$
$-\frac{1}{3}$ 0 $\frac{1}{3}$
$-\frac{1}{3}$ 0 $\frac{1}{3}$
$\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$
$\frac{1}{8}$ –1 $\frac{1}{8}$
$\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$
=
=
=
72.4
0.66
12.9
F
LP
3 MM
F
LP
3 MM
F
LP
3 MM
Average
Horizontal dark-to-light
Dark objects
Features highlighted by convolution

Fig. 3 | Example of convolution, as a core building block in deep learning in image analysis. a, Zooming in on this MRI scan of the wrist shows the individual pixels, which can be represented as a matrix of intensity values. Convolutional neural networks (CNNs) rely on learning to quantify local features in the image by multiplying the intensity values in an area (for example, a three-by-three grid) surrounding each pixel with the contents of a so-called convolution kernel, containing weights. The output image then contains the sum of these multiplications. **b**, In this example the weights are chosen such that the local feature represents the average image intensity of the area, resulting in a blurred image. **c**, Here, the weights are chosen to measure the difference in mean intensity between the left and right column, thereby highlighting horizontal transitions from dark to light. **d**, In this example the weights are chosen such that dark objects surrounded by a light background are detected.

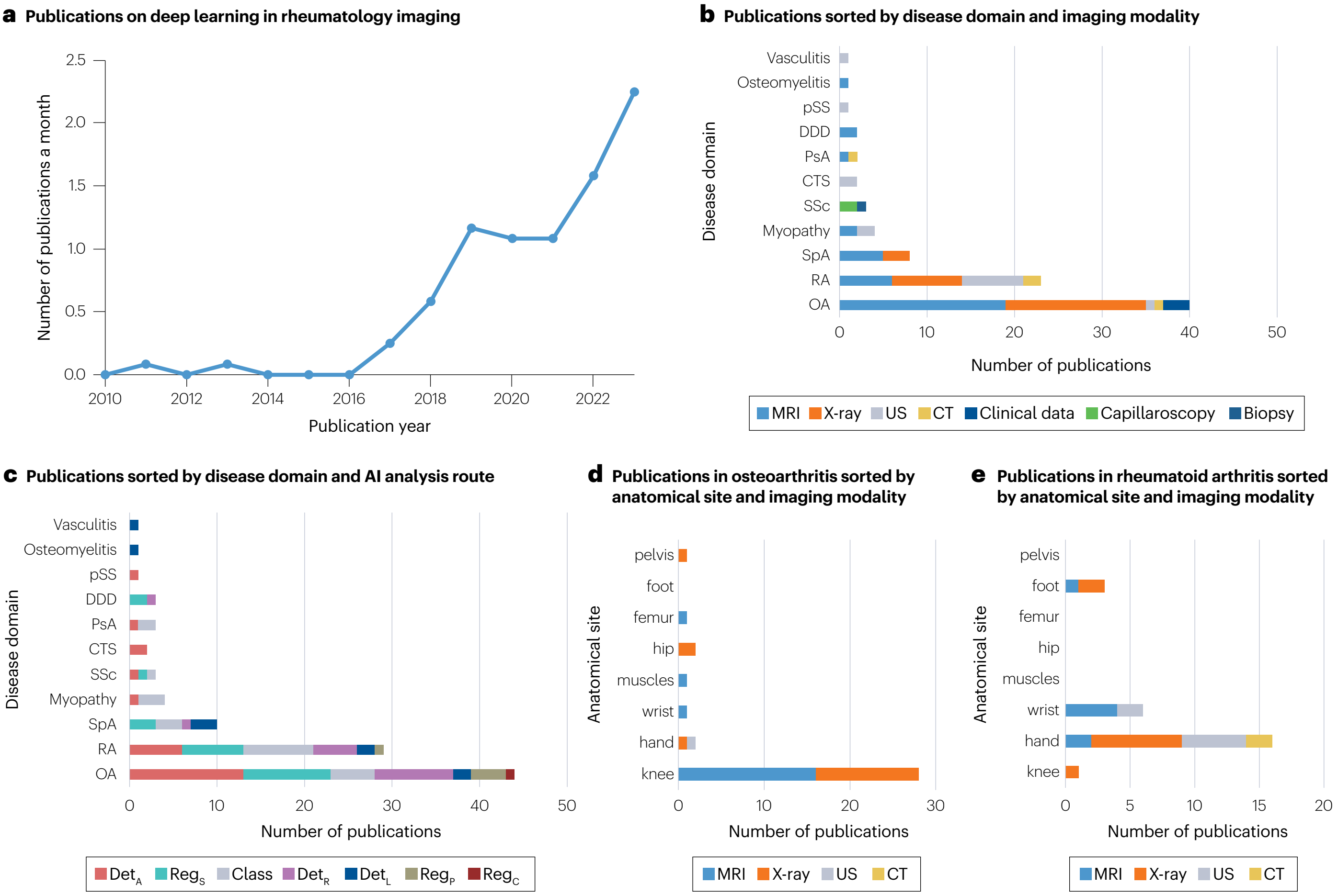

Fig. 4 | Categorization of published studies on deep learning in rheumatological imaging. a, The number of publications per month on deep learning in imaging in rheumatology. **b**, Disease domains represented in the published research, together with the use of imaging modalities in the studies within each domain. **c**, The use of deep-learning models in studies within each disease domain. **d**, The distribution of studies of deep-learning models in relation to particular anatomical sites in osteoarthritis (OA) research, along with the use of imaging modalities. **e**, The distribution of studies of deep-learning models in relation to particular anatomical sites in rheumatoid arthritis (RA) research, along with the use of imaging modalities. CTS, carpal tunnel syndrome; DDD, degenerative disc disease; Det$_A$, detection of anatomy; Det$_L$, detection of lesion or pathology; Det$_R$, detection of region of interest, slice or frame; Class, image classification; PsA, psoriatic arthritis; pSS, primary Sjögren syndrome; Reg$_C$, cross-biomarker regression; Reg$_P$, regression for prediction; Reg$_S$, regression for automatic scoring; SSc, systemic sclerosis; SpA, spondyloarthritis; US, ultrasonography.

either by detecting these lesions directly[40], or by first detecting the phalanges and subsequently classifying these ROIs based on the presence or absence of erosions[41], or by performing regression to automatically grade erosions according to the Sharp–van der Heijde method, combined with scoring joint space narrowing[42,43] and the Ratingen method to assess the severity of joint damage[44]. By taking the entire radiographic image of the hand as input and following the classification route (Fig. 2, route B), images can also be classified directly into RA or healthy[45], or into the presence or absence of subluxation and ankylosis as part of a modified total Sharp score[46].

In ultrasonography of the hand, deep learning has been applied to automatically select the most informative frame for scoring synovitis[47], similar to the frame selection for assessing the metacarpal head for OA. Given a manually or automatically selected frame, researchers have followed a two-step approach by first detecting the region of the synovium and then applying automatic grading by deep learning[48,49], or they have taken the entire frame as the input image and used a regression network to score the severity of synovitis directly[50].

In quantifying cartilage loss in RA, pilot studies have been performed using ultrasonographic images from healthy individuals to apply deep learning to detect cartilage of the distal metacarpal head and subsequently measure its thickness[51].

Exploiting the high resolution of peripheral quantitative computed tomography (CT), bone mineral density and microstructures in the bone have been quantified, by letting a deep-learning model detect the metacarpal bones and subsequently using conventional metrics to quantify bone characteristics[52]. Also using CT data as input, a neural network has been trained by taking the segmentation of the second metacarpal head and its shape as input for the classification of patients with RA or PsA, or healthy individuals[53]. This method was also applied to the classification of undifferentiated arthritis. By selecting shape features that are less dependent on image modality than image intensity features, other imaging modalities such as ultrasonography could be used to produce shape features, so that a separate neural network could be applied to classification of the different shapes[53]. The question remains, however, whether these shape parameters are sufficient to distinguish between patient groups accurately.

Using MRI scans of the hands and feet, early research was focussed on applying deep learning in the detection of carpal bones in patients with early RA[54], to serve as a basis for the quantification of bone marrow oedema (BME), erosions or synovitis. More recently, however, direct image classification was developed to distinguish between MRI scans from patients with seronegative or seropositive RA or psoriatic arthritis[7]. The multimodal approach also included clinical data, but the results indicated that the added clinical data did not provide demonstrable improvement for classification. The trained network has subsequently been applied to MRI scans of patients with psoriasis, for which further validation is still needed.

With contrast-enhanced extremity MRI of the hand and foot, deep learning has been applied to detection of the various anatomical ROIs (bones, synovia and tenosynovia) in combination with quantification (for example, tenosynovitis quantification[55]) based on the elevation of image intensity as a sign of inflammation. With longitudinal data on RA development in patients with clinically suspect arthralgia, deep-learning models were trained to perform prediction, providing a risk assessment for RA development[56]. Deep learning has also enabled exploration of treatment effects in drug evaluation trials, where the method can be used for the detection of relevant changes over time in sequential MRI scans of patients in treatment and placebo arms, without prior knowledge or hypotheses[57]. The training is in this case unsupervised (Fig. 5). By training a deep-learning network to classify each image into either a treatment or a placebo arm[8], and subsequently provide saliency maps, specific treatment effects could be localized, potentially giving a more detailed insight into the treatment mechanisms.

Deep learning in SpA

MRI of the sacroiliac joints (SIJs) has been analysed by deep learning through the detection of lesions directly[58] or in combination with the grading of BME[59] in patients with axial SpA. This route of lesion detection and quantification has also been applied to grade hip BME from MRI scans of patients with ankylosing spondylitis[60]. Through a regression network that takes the entire image as input, a method has been proposed to automate the scoring of BME, enthesitis, erosions and sclerosis in clinically suspected axial SpA[61]. By binary classification of entire MRI scans into the presence or absence of BME in the SIJ, a deep-learning network has been trained to distinguish patients with axial SpA from healthy individuals[62].

From plain radiographs, networks have been trained to grade detected ROIs of the cervical and lumbar spine according to the modified Stoke Ankylosing Spondylitis Spinal Score in patients with ankylosing spondylitis[63]. Similarly, a method has been proposed for the automatic scoring of radiographs of the SIJs, to grade radiographic sacroiliitis in patients with axial SpA[64], and neural networks have been tested for binary classification of these radiographs of the SIJ into either healthy or SpA-related sacroiliitis[65].

Deep learning in other rheumatic diseases

A variety of other applications of deep learning have been proposed, based on a range of image-acquisition modalities. These applications range from the grading of degenerative lumbar spine disorders by MRI[66] to measurement of the cross-sectional area of the median nerve from ultrasonographic images of patients with rheumatic (and musculoskeletal) diseases, to assess carpal tunnel syndrome[67]. To quantify and classify various myopathies, deep learning has been used for the analysis of MRI scans[68,69] or ultrasonographic images[70]. In patients with suspected giant-cell arteritis, colour Doppler ultrasonography images have been analysed by deep learning to automatically detect the Halo sign (hypoechoic arterial wall thickening)[71]. Finally, nailfold capillaroscopy images have been processed by deep learning to classify or quantify microvasculopathies in systemic sclerosis[72,73].

Limitations, errors and confounders

The above overview of the current literature mostly contains proposals of new methods to help analyse and assess images, which can therefore mostly be considered pilot studies. Their application in the clinic is not imminent, but it does show that AI applications should soon become clinically feasible. Reviewing these initial publications on deep learning in imaging in rheumatology shows that the research field has not yet matured, and identifies some limitations, and it also reveals that in some instances the reported accuracy might have been overestimated as a result of confounding factors and errors.

Limitations

Generalizability. The generalizability of deep-learning methods is a major topic of discussion. If a model is used 'off-label' (outside its intended use, and applied to new input data that are not represented by training, validation or testing datasets of the original study), the results might not be as accurate as in the original publication and they might give biased outcomes. Deep learning in rheumatological imaging has not yet been developed to the extent that it can detect if it is being used 'off-label', so current methods will always give an answer, whatever data are input. In the above literature overview, only very few studies[23,39,66,74] showed that the proposed model was validated on truly external test data. This observation underlines the status of most of these AI developments as initial results from pilot studies. External validation is

Glossary

Cross-entropy
A measure of the difference between two probability distributions, where the one distribution is from the ground truth and the other from the model's output. Used as a loss function, it penalizes errors especially when the model is confident, but wrong.

Data augmentation
Artificially expanding the number and diversity of training examples by performing random transformations, or adding noise or simulated objects (such as lesions) to existing image data.

Dice coefficient
A statistic that is used to quantify the similarity between two samples; in image segmentation, it measures the overlap between the ground truth and the model-produced segmentation[3].

Loss function
A mathematical entity for quantifying how well a machine-learning algorithm models the data. Higher values indicate poorer modelling ability. During training, the loss function is coupled with an optimizer, which is used to tune the parameters of the machine-learning or deep-learning algorithm to minimize the loss function and ultimately maximize algorithm performance[3].

Saliency maps
Derived images usually displayed as heat maps that show the locations in the input image that contributed most to the model's output.

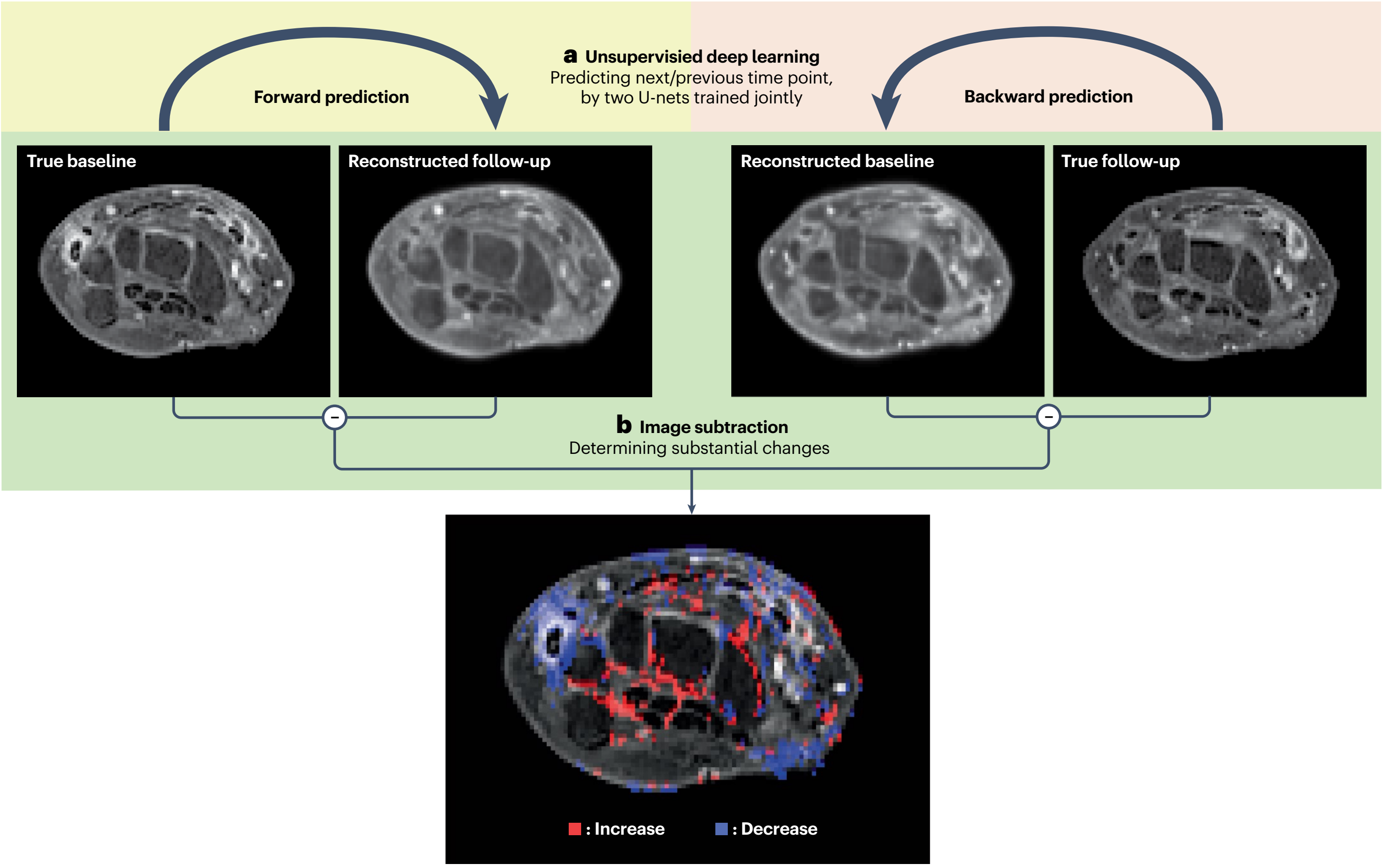

Fig. 5 | AI-based comparison between baseline and follow-up MRI of the wrist, without prior knowledge. In this example, the neural network has learned without supervision to neglect changes that are common in the population (for example, common artefacts or gradual changes in the MRI scanner) and to only show intensity changes that cannot be predicted[87]. **a**, The follow-up image is predicted from the baseline image, and vice versa. Two so-called U-nets are trained jointly with a common loss function, in which the input image at one time point is encoded followed by decoding into the predicted image at the opposite time point. **b**, By subtracting the reconstructed images from the real images and combining the positive parts of these different images, the unpredictable (and therefore uncommon) changes over time are detected and colour-coded in the change map. This approach can be considered a 'detection' pathway, where in this case a series of images are used as input instead of a single image. The change maps can be used to visually assess or score inflammatory changes by pinpointing relevant changes, or they can be used as input again for a second deep-learning network to classify or predict patient outcomes.

usually performed only when a deep-learning model becomes (part of) a commercial product.

Even after external validation, generalizability remains a concern, because new imaging machines and image-acquisition protocols will continue to be developed and put onto the market. For example, MRI scanners with novel sequences will produce a new appearance of lesions and normal anatomy, for which even human experts need to be retrained. Currently, it is not plausible that deep learning will be less sensitive to these technical variations than humans.

Dependency on large datasets. Another limitation of current deep-learning techniques is their dependency on large datasets for training. A striking difference in training efficiency between humans and AI still exists. Where humans only need a small set of image data to understand anatomy, and to directly apply this knowledge in a broader context, AI still needs large numbers of strictly defined input images with clearly defined gold-standard output data. In AI applications that involve general photography, a large amount of data is available through the internet, whereas medical imaging datasets are limited, not only because of images being produced less frequently than, for example, imagery of cats and dogs, but also because of ethical and privacy concerns in relation to medical imaging.

A minimum amount of image data required for a deep-learning model is difficult to define, as it depends on the size of the model and the problem at hand. Generally, more training data are required for image classification and regression than for image segmentation, because the latter contains many repeated examples located within an image. This difference is reflected in the size of the training datasets in published studies; for lesion or anatomy detection tasks, typically hundreds of images are used, whereas for regression tasks thousands of images are required.

As a consequence of these relatively small datasets, the risk of a model becoming over-fitted to the data is considerable. Data augmentation can help to overcome this issue, and synthetic image generation can also lower these risks. However, larger original datasets with high-quality gold standards (for example, manually annotated lesions or visual scores) are needed that represent the ground truth accurately. The OA Initiative is a good example of how deep-learning research can be boosted by the availability of public datasets together with gold-standard and clinical data. However, these public datasets do intrinsically tend to promote a certain research direction. If, for example, only image data with gold-standard image segmentation are provided, then AI researchers will be motivated to only follow the route focussing on the detection of anatomy or lesions, thereby missing out on potential benefits from following the classification and/or regression routes. Other concerns with public datasets are that it might be difficult to check the quality of image data and the provided gold standards, and to check whether the same schema for annotation is used that is relevant for all studies that use these data. However, the advantages of these public datasets clearly outweigh the disadvantages, and more publicly available data are required for all image modalities and disease domains.

Explainability. Explaining the decisions made by deep-learning models requires the opening of the black box via the use of saliency maps. These maps can provide a 'sanity check' for the training process: if a map shows hotspots that are completely outside any ROI, it indicates that the model has been trained incorrectly, or has learned an irrelevant but deterministic shortcut. Even when a saliency map points to realistic signs in an image, it should be noted that humans are interpreting these maps and imposing their expectations on them. We should be careful not to only accept the quality of a saliency map if it meets our expectations or to select a mapping method only based on its visual appeal[75]. Saliency-mapping techniques are still at an early developmental stage, which means that any unexpected finding could represent an error in the deep-learning model but it could also indicate an error in the mapping method. Notably, validation is complicated when we don't know what a saliency map should actually look like. For day-to-day photography of cats and dogs it is usually very clear to humans that the saliency map should point to the cat or dog. For regression tasks, where AI has learned to predict pain indices, for example, it is far more difficult to verify that the saliency maps are realistic. The fact that saliency maps were presented in only 30% of the studies assessed for this Review limits the explainability of the majority of the presented models, as it is unclear whether they would pass the sanity check.

Image reconstruction. Although image reconstruction has not been considered an element of automated image interpretation, it actually does have an important role in the quality of AI interpretation. Image reconstruction has traditionally been optimized for human visual interpretation, whereas AI could have different requirements. As illustrated in Fig. 1, the human visual system can lack sensitivity to changes or differences in absolute image intensities, so there is not necessarily a strict requirement for accurately calibrated images for qualitative interpretation by humans. By contrast, even slight deviations in image-intensity ranges can have devastating effects on AI-based quantification (as described in the 'Batch effect' section below). Therefore, especially for deep-learning interpretation, the parameters of image reconstruction and/or enhancement in MRI, ultrasonography, plain radiography and CT should be reconsidered and optimized to take full advantage of the whole image-acquisition and post-processing chain. As deep learning has also entered the domain of image reconstruction, there is an opportunity to perform so-called end-to-end training, where the weights, biases and hyper-parameters of both reconstruction and post-processing are optimized simultaneously.

Using prior knowledge or hypothesis-free analysis? Most papers on deep learning in rheumatological image interpretation describe studies that followed the detection–quantification route, in which lesions or anatomical structures (or at least ROIs) are detected, followed by quantification or automatic scoring within those areas. This approach intrinsically assumes that the relevant information should only come from these predefined areas in the image. This assumption is usually based on prior knowledge collected over decades of human experience, which might be completely justifiable in most situations. In some specific instances, however, such as in assumptions of the absence of anatomical structures (for example, sheaths surrounding extensor and flexor tendons crossing metatarsophalangeal joints[76]) and, therefore, the assumed absence of any relevant inflammatory signs in these structures, the assumptions proved to be false. In deep learning such false assumptions can be prevented by taking the entire image as input for classification or grading by regression. The accompanying saliency maps could then point to areas where relevant signs are found, without any prior assumptions or hypotheses, which might differ from the expected areas. Current deep-learning approaches, however, are still using the conventional approach of AI-based segmentation and quantification, which unnecessarily limits the receptive field and potentially conceals relevant information outside these predefined anatomical areas.

Confounders and errors

Errors in training and validation design can result in considerable over-estimation of the performance of deep-learning methods[77]. Therefore, a user of AI software should at least be aware of possible confounders and errors, before relying on the accuracy of the results (Box 1). Although it is not possible to provide an exhaustive list, some known confounders and errors in deep-learning approaches are discussed below.

Data leakage. Many of the errors and confounders in deep-learning studies can be traced back to the design of the training and validation. In studies involving repeated scans of patients, or images that are subdivided into separate slices or ROIs, a phenomenon called 'data leakage' can potentially occur. If the data are subdivided into training, validation and testing datasets at any level below the patient level, the same source of data can occur in any of these datasets. Therefore, during validation or testing, the deep-learning method could be presented with imaging data that it has already 'seen' during training. As a result, the validation and testing results can be over-estimated. In a number of the studies considered for this Review, data leakage could not be ruled out because it was not mentioned explicitly at which level repeated data were split into separate datasets.

Batch effect. Even without data leakage, training data can contain a dataset bias, which can occur, for example, when one patient group is (predominantly) collected from a different hospital or via a different model or brand of an imaging modality than another patient group. Especially when the task is to classify these two patient groups or estimate any variable that can be deduced from the patient groups, this bias can result in a situation where the model has learned irrelevant information that is still distinctive or predictive for the target but that was not an intended learning outcome. In this example, any sign in the image that indicates which imaging system was used can inform the disease classification[78], leading to a phenomenon called shortcut learning[79]. By external validation and by using saliency maps, shortcut learning can be revealed.

Box 1

Checklist of requirements that a deep-learning model must meet, before considering applying it in clinical research

Has the method been properly designed and validated?

- Validated with an external test set?
- No confounders or errors occurred, such as data leakage and batch effects?
- Have imbalanced data been compensated for?
- In supervised learning, is the quality of the gold standard high enough?
- Has an evaluation metric been used that is relevant to our application?

Do the saliency maps make sense, so there is no shortcut learning?

Is our patient population represented by the data used during development, training and validation?

- If the data are imbalanced, does our patient population fall within the infrequent category or subgroup?

Is our imaging modality compatible with those of the testing set and external test set?

Is our image acquisition protocol similar to those used during development and validation?

Is the ground truth used for training the same as the outcome we need?

- Is it created with a schema that matches our needs?
- Has the schema been applied consistently?

For clinical use, FDA and EMA regulations will be applicable, but for clinical research a critical view remains important. A more general checklist for AI papers has been published as a guide for authors and reviewers[88].

Imbalanced data. If a patient characteristic such as disease severity has a skewed distribution, or if classes are not evenly distributed, it can occur that during training many of the batches contain data from only one subgroup or class. This occurrence will hamper the training process, especially when, for example, only accuracy is used as a performance metric. Balancing the dataset by removing a large part of the majority subgroup or class is often not an option, as medical data are already sparse. Several technical solutions are available using the whole dataset, for example, by under-sampling the majority subgroup or class for each batch, putting more weight on the minority subgroup or class or increasing data augmentation. The fact remains, however, that the model has fewer examples from the under-represented subgroup or class to learn from, which might have an effect on the performance when the model is applied in practice where this balance is different.

Reuse of test data. Ideally, the test dataset is touched only once, after the final model has been chosen and its hyper-parameters optimized and fully trained, solely using training and validation data. However, it is possible that a trivial error can emerge at the final stage of testing, at which point it is tempting to redesign, retrain and retest the model, and publish only this latest version. In such an instance, the test set has strictly speaking been used for training. In 'grand challenges' the same reuse of test data can occur when a research group re-enters new models to the dashboard, thereby indirectly learning from the provided test set. They will then climb the scoreboard, not necessarily with a more accurate model.

Overfitting. The problem of overfitting occurs when a deep-learning model performs accurately on the training dataset, but less so on the test set. The large number of weights and biases in deep-learning networks in combination with the limited availability of training sets increases the risk of overfitting, in which the model fits only to the specific training dataset. If researchers do not use an independent test set, sampled from the same distribution as the training set, overfitting can go unnoticed.

Cherry picking from random initializations. In many models the weights and biases are initialized by a method that uses random numbers (from random seeds), before starting the training process. The results after different training sessions can therefore vary. During validation, it is then tempting to only present the results from the 'best' initialization, but to show the robustness of the model, the average and standard deviation over a range of training sessions should be presented. As with overfitting, this cherry picking can lead to disappointing results during the testing phase, where the weights and biases need to be fixed.

Performance metrics. During training, the performance of a neural network is monitored by a loss function, in order to reach an optimum. The choice of loss function is determined by the task at hand, usually cross-entropy and Dice coefficient for detection, cross-entropy for classification and mean-squared error for regression. The same holds true for the final validation of the trained network, where a proper choice of performance metric is needed. Generally, this performance metric is the same as the loss function used for training, and auxiliary performance metrics are used to give a full overview of all aspects of the performance. The use of inappropriate performance metrics or the use of performance metrics inappropriately can result in the reporting of misleading results[80].

Errors in ground truth data. Specifically in detection tasks, the quality of the training depends on the quality of the ground-truth segmentations, which are generated by human experts. The results can be compromised by the variation in annotation schemas. For example, in determining how lesions or anatomical structures are defined and implemented in a protocol, there is the question of whether to include or exclude erosions when annotating bones. Variation can also result from inter-observer and intra-observer variability, systematic differences between observers or learning-curve effects during annotation, all of which can also create batch effects[81].

Errors in implementation. The software implementation of a model and its training strategy can contain software bugs that affect the validity of the results. These errors in computer programming are virtually impossible to catch from analysis of published data. Only publishing the computer code along with the article can reduce this type of risk, although it takes considerable effort for reviewers to also check this code. External validation by other research groups using the same trained network might enable identification of this type of error in already published studies.

Incorrect interpretation of saliency maps. It is tempting to interpret saliency maps as the results of lesion detection by the deep-learning network. However, this assumption is not always justifiable. For example, if a task is to grade disease severity solely based on image data, the assumption might indeed be valid. However, if the task is to distinguish between two populations, the saliency map might just show areas where the populations mostly differ, not necessarily involving lesions, which might occur in both groups.

General confounders. Apart from confounders resulting from poor training and validation design, common confounders that generally occur in research can also occur during deep-learning validation. As a hypothetical example, any difference between genders can be picked up by deep-learning algorithms and used for prediction of the occurrence of rheumatic diseases; it is easy for a neural network to select a feature that is related to the size of the patient (and therefore to gender) by simply counting pixels. This confounding can therefore lead to shortcut learning, where a risk factor is quantified instead of a patient outcome.

Consequences for rheumatologists and radiologists

Once the possible confounders and errors are identified and resolved, how might AI influence the daily work of clinicians in the near future? Technological innovations and AI in particular are subject to high expectations. Health care is becoming increasingly expensive, and many health care systems face potential staffing shortfalls. This dual challenge calls for radical solutions, for example, through increased use of medical technology, and in this regard, AI is believed to have great potential. Here we outline some possible consequences of the use of AI for physicians, working within or closely related to rheumatology.

With respect to images and image interpretation, deep learning could partly replace the radiologist by either providing a fully automated written image result, or by indicating the areas of potential interest and/or abnormalities that should subsequently be evaluated by the human eye. In this way the time spent by radiologists in image evaluation could be reduced. Although this outcome has not yet occurred, commercial AI products are now becoming available, for example, for knee OA grading (a list of commercial AI products for radiology can be found at grand-challenge.org). An important consideration for the future implementation of these AI products is whether they add value to the daily workflow and to image interpretation. These products also need to be compatible with the hospital's archiving and communication systems.

Rheumatologists have a slightly different task in the diagnostic process to radiologists, because rheumatologists combine the results of radiological imaging with clinical data from medical history, physical examination and laboratory results. The diagnostic process for rheumatologists relies on human pattern recognition developed during years of training. Expertise or 'gut feeling' has a role in this diagnostic process. As pattern finding is inherent in machine learning, it is to be expected that deep learning could also be useful here. Models could be trained by the inclusion not only of images but also of comorbidities, symptoms, signs, medication use and laboratory results (multimodal deep learning).

Training appropriate AI tools will require enormous datasets and efforts. Furthermore, regulatory systems need to be able to handle technological issues with new AI applications that might not be easily verifiable by human experts. For example, AI-based prediction of disease progression or therapeutic response or production of saliency maps requires a different type of verification from AI-based detection systems, because of differences in ground truth data. As mentioned in the context of saliency maps, explainability and traceability are important for the acceptance and adoption of deep learning models as medical devices. Moreover, quality assurance procedures need to be adjusted, to guarantee stability of a model's performance, with a strong link to the (stability of) image-acquisition systems and protocols used, and with reference to the intended clinical use. In this Review, we mainly focussed on the scientific and technological challenges of developing and applying deep learning in rheumatological imaging, but legal and ethical concerns will also need to be addressed. Notably, these concerns are general and also apply to non-imaging-based AI applications, that are designed to support medical decision-making and workflow and to improve cost-effectiveness[82]. Additionally, regulations apply for safety, the ethical use of sensitive image information and protection of privacy, and these regulations vary across jurisdictions[83], which represents a challenge for software companies' business models.

If successfully trained, fully validated, regulated and quality-assured, deep learning could be a valuable asset for diagnosis and treatment decision-making processes in rheumatology practices that could reduce the workload of radiologists and rheumatologists working in secondary care.

AI, combined with (patient-friendly) imaging, might also provide opportunities to change health care in the setting between primary and secondary care. Many rheumatological diagnoses have clinical arthritis as a crucial feature. Assessment of joint inflammation for this arthritis involves physical examination by rheumatologists. Gaining experience takes years of training, and general practitioners are often unable to detect clinical arthritis at an early stage. Consequently, guidelines recommend early referral for suspected clinical arthritis. Subsequently, at secondary care assessment, a high proportion of referred patients do not have clinical arthritis, so their referrals can be considered unnecessary. The use of other methodologies to detect clinical arthritis could solve this problem. Several referral tools (questionnaires) exist, but none is an accurate proxy of joint inflammation, as assessed at joint examination by rheumatological experts. Imaging (in particular MRI) is sensitive for the detection of joint inflammation and could be useful to this end. High sensitivity is intrinsic to MRI, and high specificity can be obtained by comparison with MRI results from healthy individuals matched for age, gender and joint location. This comparison is required because, in some synovia and bones, subtle changes resembling inflammation (synovitis, BME) are normally present, especially in older age[84,85]. Although MRI is currently considered expensive and unfeasible for screening before secondary care, this situation could change with the ongoing development of short MRI sequences that do not require contrast enhancement[86], which could prove to be valuable, accurate and cost-effective for the early detection of joint inflammation. Ultimately, a brief MRI scan of (for instance) hand joints, followed by rapid AI-based analysis, could be valuable for the identification of patients who require early access to a rheumatologist.

In the settings mentioned above, AI is trained (supervised) in a one-way process, with clinical evaluation as the reference. However, a more complex arrangement could also be possible, with interactions back and forth between the computer and the clinician that could possibly improve understanding of the disease pathology underlying joint inflammation and induce a refinement of expertise in clinical practice.

Achieving the conceptual advances described above not only requires a great deal of work by engineers and clinicians for development and evaluation, but also raises ethical issues. Opportunities might come with the following risks, which should be carefully considered and monitored. The role of software companies should be scrutinized, with assessment of their business models and of whether they are moving the field forward in a cost-effective way or looking for maximum financial gain. A comparison can be made with pharmaceutical companies that market newly developed drugs at high prices that are nonetheless accepted by health care providers on the basis of the associated clinical benefits. A discussion on the costs and benefits of AI techniques is necessary, and it is hoped that financial gain for companies will not hinder the availability of these techniques wherever they have the potential to improve patient outcomes.

In summary, deep learning from radiological images, whether or not it is combined with other clinical data, has enormous potential to support the work of clinicians in secondary care and possibly also before referral to secondary care. However, in addition to technical challenges, this promise also requires simultaneous discussion of related practical, regulatory, ethical and financial issues.

Conclusions

The literature relating to deep learning in rheumatological imaging is beginning to grow exponentially, with applications in a wide range of rheumatic diseases (in particular OA, RA and SpA). Most publications so far are associated with pilot studies that commonly lack external validation. However, commercial products with deep-learning solutions that are extensively validated are also beginning to emerge.

A critical view of the proposed methods is valuable as it can help to identify the confounders and errors that can lead to overestimation of the accuracy of deep-learning tools. Most of the methods considered in this Review are based on supervised learning, which requires large numbers of images and high-quality ground truth data. Despite these notes of caution, it is clear that if carefully validated, deep learning will be able to help rheumatologists and musculoskeletal radiologists to perform their diagnostic, prognostic and monitoring tasks.

By shifting from the paradigm of detecting objects (with quantification and classification) to image classification or regression, deep learning could induce a transition from machine learning towards machine teaching, in which images are analysed without prior hypotheses and the resultant saliency maps can produce new information. By this process, AI could detect additional signs of disease that are not currently discernible by human experts. With the use of cross-biomarker regression, this approach can be extended further by letting a neural network locate areas in an image that are predominantly associated with symptoms, overall function or treatment response. Other opportunities for applications involving deep learning include prediction of the development of rheumatic diseases from baseline images, and prediction of treatment response for personalized treatment strategies. With the rapid progress that is occurring in the field of deep-learning methods, it is likely that we cannot yet foresee all of their possible applications in rheumatological imaging.

Published online: 08 February 2024

References

1. Kingsmore, K. M., Puglisi, C. E., Grammer, A. C. & Lipsky, P. E. An introduction to machine learning and analysis of its use in rheumatic diseases. *Nat. Rev. Rheumatol.* **17**, 710–730 (2021).
2. Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
3. Calivà, F. et al. Studying osteoarthritis with artificial intelligence applied to magnetic resonance imaging. *Nat. Rev. Rheumatol.* **18**, 112–121 (2022).
4. Cipolletta, E. et al. Artificial intelligence for ultrasound informative image selection of metacarpal head cartilage. a pilot study. *Front. Med.* **8**, 589197 (2021).
5. Prasoon, A. et al. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med. Image Comput. Comput. Assist. Interv.* **16**, 246–253 (2013).
6. Banerjee, S., Bhunia, S. & Schaefer, G. Osteophyte detection for hand osteoarthritis identification in X-ray images using CNNs. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2011**, 6196–6199 (2011).
7. Folle, L. et al. Advanced neural networks for classification of MRI in psoriatic arthritis, seronegative, and seropositive rheumatoid arthritis. *Rheumatology* **61**, 4945–4951 (2022).
8. Hassanzadeh, T. et al. AB0205 RA treatment effects in wrist MRIs, determined by deep learning. *Ann. Rheum. Dis.* **82**, 1286 (2023).
9. Abedin, J. et al. Predicting knee osteoarthritis severity: comparative modeling based on patient's data and plain X-ray images. *Sci. Rep.* **9**, 5761 (2019).
10. Leung, K. et al. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. *Radiology* **296**, 584–593 (2020).
11. Jia, J. et al. Automatic pulmonary function estimation from chest CT scans using deep regression neural networks: the relation between structure and function in systemic sclerosis. *IEEE Access* **11**, 135272–135282 (2023).
12. Chang, G. H. et al. Assessment of knee pain from MR imaging using a convolutional Siamese network. *Eur. Radiol.* **30**, 3538–3548 (2020).
13. Ras, G., Xie, N., Gerven, M. V. & Doran, D. Explainable deep learning: a field guide for the uninitiated. *J. Artif. Int. Res.* **73**, 68 (2022).
14. National Institutes of Health. The Osteoarthritis Initiative. *NIMH Data Archive* https://nda.nih.gov/oai (2023).
15. Chen, N. et al. A fully automatic target detection and quantification strategy based on object detection convolutional neural network YOLOv3 for one-step X-ray image grading. *Anal. Methods* **15**, 164–170 (2023).
16. Chen, P., Gao, L., Shi, X., Allen, K. & Yang, L. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Comput. Med. Imaging Graph.* **75**, 84–92 (2019).
17. Liu, B., Luo, J. & Huang, H. Toward automatic quantification of knee osteoarthritis severity using improved Faster R-CNN. *Int. J. Comput. Assist. Radiol. Surg.* **15**, 457–466 (2020).
18. Norman, B., Pedoia, V., Noworolski, A., Link, T. M. & Majumdar, S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J. Digit. Imaging* **32**, 471–477 (2019).
19. Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P. & Saarakkala, S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. *Sci. Rep.* **8**, 1727 (2018).
20. Tolpadi, A. A., Lee, J. J., Pedoia, V. & Majumdar, S. Deep learning predicts total knee replacement from magnetic resonance images. *Sci. Rep.* **10**, 6371 (2020).
21. Tiulpin, A. et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *Sci. Rep.* **9**, 20038 (2019).
22. Hirvasniemi, J. et al. The KNee OsteoArthritis Prediction (KNOAP2020) challenge: an image analysis challenge to predict incident symptomatic radiographic knee osteoarthritis from MRI and X-ray images. *Osteoarthritis Cartilage* **31**, 115–125 (2023).
23. Jansen, M. P. et al. Artificial intelligence in osteoarthritis: repair by knee joint distraction shows association of pain, radiographic and immunological outcomes. *Rheumatology* **62**, 2789–2796 (2022).
24. Ambellan, F., Tack, A., Ehlke, M. & Zachow, S. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: data from the Osteoarthritis Initiative. *Med. Image Anal.* **52**, 109–118 (2019).
25. Cheng, R. et al. Fully automated patellofemoral MRI segmentation using holistically nested networks: implications for evaluating patellofemoral osteoarthritis, pain, injury, pathology, and adolescent development. *Magn. Reson. Med.* **83**, 139–153 (2020).
26. Gaj, S., Yang, M., Nakamura, K. & Li, X. Automated cartilage and meniscus segmentation of knee MRI with conditional generative adversarial networks. *Magn. Reson. Med.* **84**, 437–449 (2020).
27. Liu, F. et al. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. *Magn. Reson. Med.* **79**, 2379–2391 (2018).
28. Norman, B., Pedoia, V. & Majumdar, S. Use of 2D U-net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. *Radiology* **288**, 177–185 (2018).
29. Panfilov, E., Tiulpin, A., Nieminen, M. T., Saarakkala, S. & Casula, V. Deep learning-based segmentation of knee MRI for fully automatic subregional morphological assessment of cartilage tissues: data from the Osteoarthritis Initiative. *J. Orthop. Res.* **40**, 1113–1124 (2022).

30. Razmjoo, A. et al. T_2 analysis of the entire osteoarthritis initiative dataset. *J. Orthop. Res.* **39**, 74–85 (2021).
31. Chang, G. H. et al. Subchondral bone length in knee osteoarthritis: a deep learning-derived imaging measure and its association with radiographic and clinical outcomes. *Arthritis Rheumatol.* **73**, 2240–2248 (2021).
32. Pedoia, V. et al. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. *J. Magn. Reson. Imaging* **49**, 400–410 (2019).
33. Liu, F. et al. Deep learning approach for evaluating knee MR images: achieving high diagnostic performance for cartilage lesion detection. *Radiology* **289**, 160–169 (2018).
34. Astuto, B. et al. Automatic deep learning-assisted detection and grading of abnormalities in knee MRI studies. *Radiol. Artif. Intell.* **3**, e200165 (2021).
35. Namiri, N. K. et al. Deep learning for large scale MRI-based morphological phenotyping of osteoarthritis. *Sci. Rep.* **11**, 10915 (2021).
36. Brui, E. et al. Deep learning-based fully automatic segmentation of wrist cartilage in MR images. *NMR Biomed.* **33**, e4320 (2020).
37. Üreten, K. et al. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. *Skeletal Radiol.* **49**, 1369–1374 (2020).
38. 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* **12**, e0178992 (2017).
39. von Schacky, C. E. et al. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. *Radiology* **295**, 136–145 (2020).
40. Radke, K. L. et al. Adaptive IoU thresholding for improving small object detection: a proof-of-concept study of hand erosions classification of patients with rheumatic arthritis on X-ray images. *Diagnostics* **13**, 104 (2022).
41. Murakami, S., Hatano, K., Tan, J., Kim, H. & Aoki, T. Automatic identification of bone erosions in rheumatoid arthritis from hand radiographs based on deep convolutional neural network. *Multimed. Tools Appl.* **77**, 10921–10937 (2018).
42. Hirano, T. et al. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. *Rheumatol. Adv. Prac.* **3**, rkz047 (2019).
43. Chaturvedi, N. DeepRA: predicting joint damage from radiographs using CNN with attention. Preprint at https://doi.org/10.48550/arXiv.2102.06982 (2021).
44. Rohrbach, J., Reinhard, T., Sick, B. & Dürr, O. Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks. *Comput, Electr. Eng.* **78**, 472–481 (2019).
45. Üreten, K., Erbay, H. & Maraş, H. H. Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. *Clin. Rheumatol.* **39**, 969–974 (2020).
46. Izumi, K. et al. Detecting hand joint ankylosis and subluxation in radiographic images using deep learning: a step in the development of an automatic radiographic scoring system for joint destruction. *PLoS ONE* **18**, e0281088 (2023).
47. Fiorentino, M. C., Moccia, S., Cipolletta, E., Filippucci, E. & Frontoni, E. A learning approach for informative-frame selection in US rheumatology images. in: Cristani, M., Prati, A., Lanz, O., Messelodi, S., Sebe, N. (eds) New Trends in Image Analysis and Processing — ICIAP 2019. ICIAP 2019. Lecture Notes in Computer Science, vol 11808. https://doi.org/10.1007/978-3-030-30754-7_23 (Springer, Cham, 2019).
48. Tang, J. et al. Enhancing convolutional neural network scheme for rheumatoid arthritis grading with limited clinical data. *Chin. Phys. B* **28**, 038701 (2019).
49. Hemalatha, R. J., Vijaybaskar, V. & Thamizhvani, T. R. Automatic localization of anatomical regions in medical ultrasound images of rheumatoid arthritis using deep learning. *Proc. Inst. Mech. Eng. H.* **233**, 657–667 (2019).
50. Christensen, A. B. H., Just, S. A., Andersen, J. K. H. & Savarimuthu, T. R. Applying cascaded convolutional neural network design further enhances automatic scoring of arthritis disease activity on ultrasound images from rheumatoid arthritis patients. *Ann. Rheum. Dis.* **79**, 1189–1193 (2020).
51. Fiorentino, M. C. et al. A deep-learning framework for metacarpal-head cartilage-thickness estimation in ultrasound rheumatological images. *Comput. Biol. Med.* **141**, 105117 (2022).
52. Folle, L. et al. Deep learning methods allow fully automated segmentation of metacarpal bones to quantify volumetric bone mineral density. *Sci. Rep.* **11**, 9697 (2021).
53. Folle, L. et al. Deep learning-based classification of inflammatory arthritis by identification of joint shape patterns — how neural networks can tell us where to "Deep Dive" clinically. *Front. Med.* **9**, 850552 (2022).
54. Wong, L. M., Shi, L., Xiao, F. & Griffith, J. F. Fully automated segmentation of wrist bones on T2-weighted fat-suppressed MR images in early rheumatoid arthritis. *Quant. Imaging Med. Surg.* **9**, 579–589 (2019).
55. Shamonin, D. P. et al. POS0920 quantification of tenosynovitis from wrist MRIs, based on deep learning. *Ann. Rheum. Dis.* **82**, 770–771 (2023).
56. Li, Y. et al. OP0002 exploring the use of artificial intelligence in predicting rheumatoid arthritis, based on extremity MR scans in early arthritis and clinically suspect arthralgia patients. *Ann. Rheum. Dis.* **82**, 1–2 (2023).
57. Hassanzadeh, T. et al. A deep learning-based comparative MRI model to detect inflammatory changes in rheumatoid arthritis. *Biomed. Signal. Process. Control.* **88**, 105612 (2024).
58. Hepburn, C. E. et al. Towards deep learning-assisted quantification of inflammation in spondyloarthritis: intensity-based lesion segmentation. Preprint at https://doi.org/10.48550/arXiv.2106.11343 (2021).
59. Lin, K. Y. Y., Peng, C., Lee, K. H., Chan, S. C. W. & Chung, H. Y. Deep learning algorithms for magnetic resonance imaging of inflammatory sacroiliitis in axial spondyloarthritis. *Rheumatology* **61**, 4198–4206 (2022).
60. Han, Q. et al. Automatic quantification and grading of hip bone marrow oedema in ankylosing spondylitis based on deep learning. *Mod. Rheumatol.* **32**, 968–973 (2022).
61. Bressem, K. K. et al. Deep learning detects changes indicative of axial spondyloarthritis at MRI of sacroiliac joints. *Radiology* **305**, 655–665 (2022).
62. Lee, K. H., Choi, S. T., Lee, G. Y., Ha, Y. J. & Choi, S. I. Method for diagnosing the bone marrow edema of sacroiliac joint in patients with axial spondyloarthritis using magnetic resonance image analysis based on deep learning. *Diagnostics* **11**, 1156 (2021).
63. Koo, B. S. et al. A pilot study on deep learning-based grading of corners of vertebral bodies for assessment of radiographic progression in patients with ankylosing spondylitis. *Ther. Adv. Musculoskelet. Dis.* **14**, 1759720x221114097 (2022).
64. Bressem, K. K. et al. Deep learning for detection of radiographic sacroiliitis: achieving expert-level performance. *Arthritis Res. Ther.* **23**, 106 (2021).
65. Üreten, K., Maraş, Y., Duran, S. & Gök, K. Deep learning methods in the diagnosis of sacroiliitis from plain pelvic radiographs. *Mod. Rheumatol.* **33**, 202–206 (2023).
66. Grob, A. et al. External validation of the deep learning system "SpineNet" for grading radiological features of degeneration on MRIs of the lumbar spine. *Eur. Spine J.* **31**, 2137–2148 (2022).
67. Smerilli, G. et al. Development of a convolutional neural network for the identification and the measurement of the median nerve on ultrasound images acquired at carpal tunnel level. *Arthritis Res. Ther.* **24**, 38 (2022).
68. Fabry, V. et al. A deep learning tool without muscle-by-muscle grading to differentiate myositis from facio-scapulo-humeral dystrophy using MRI. *Diagn. Interv. Imaging* **103**, 353–359 (2022).
69. Wang, F. et al. Assessment of idiopathic inflammatory myopathy using a deep learning method for muscle T2 mapping segmentation. *Eur. Radiol.* **33**, 2350–2357 (2022).
70. Burlina, P., Billings, S., Joshi, N. & Albayda, J. Automated diagnosis of myositis from muscle ultrasound: exploring the use of machine learning and deep learning methods. *PLoS ONE* **12**, e0184059 (2017).
71. Roncato, C. et al. Colour Doppler ultrasound of temporal arteries for the diagnosis of giant cell arteritis: a multicentre deep learning study. *Clin. Exp. Rheumatol.* **38**, 120–125 (2020).
72. Garaiman, A. et al. Vision transformer assisting rheumatologists in screening for capillaroscopy changes in systemic sclerosis: an artificial intelligence model. *Rheumatology* **62**, 2492–2500 (2022).
73. Gurunath Bharathi, P. et al. A deep learning system for quantitative assessment of microvascular abnormalities in nailfold capillary images. *Rheumatology* **62**, 2325–2329 (2023).
74. Mohajer, B. et al. Role of thigh muscle changes in knee osteoarthritis outcomes: osteoarthritis initiative data. *Radiology* **305**, 169–178 (2022).
75. Adebayo, J. et al. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31* (NeurIPS, 2018).
76. Dakkak, Y. J., Jansen, F. P., DeRuiter, M. C., Reijnierse, M. & van der Helm-van Mil, A. H. M. Rheumatoid arthritis and tenosynovitis at the metatarsophalangeal joints: an anatomic and MRI study of the forefoot tendon sheaths. *Radiology* **295**, 146–154 (2020).
77. Maleki, F. et al. Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. *Radiol. Artif. Intell.* **5**, e220028 (2023).
78. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
79. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
80. Reinke, A. et al. Understanding metric-related pitfalls in image analysis validation. Preprint at https://doi.org/10.48550/arXiv.2302.01790 (2023).
81. Abdalla, M. & Fine, B. Hurdles to artificial intelligence deployment: noise in schemas and "Gold" labels. *Radiol. Artif. Intell.* **5**, e220056 (2023).
82. Meszaros, J., Minari, J. & Huys, I. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Front. Genet.* **13**, 927721 (2022).
83. Pesapane, F., Volonté, C., Codari, M. & Sardanelli, F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* **9**, 745–753 (2018).
84. Mangnus, L., van Steenbergen, H. W., Reijnierse, M. & van der Helm-van Mil, A. H. Magnetic resonance imaging-detected features of inflammation and erosions in symptom-free persons from the general population. *Arthritis Rheumatol.* **68**, 2593–2602 (2016).
85. Boer, A. C. et al. Using a reference when defining an abnormal MRI reduces false-positive MRI results-a longitudinal study in two cohorts at risk for rheumatoid arthritis. *Rheumatology* **56**, 1700–1706 (2017).
86. Boeren, A. M. P. et al. Towards a simplified fluid-sensitive MRI protocol in small joints of the hand in early arthritis patients: reliability between modified Dixon and regular Gadolinium enhanced TSE fat saturated MRI-sequences. *Skeletal Radiol.* **52**, 1193–1202 (2023).
87. Hassanzadeh, T. et al. A deep learning model to locate inflammatory changes in rheumatoid arthritis. *Ann. Rheum. Dis.* **82**, 298–299 (2023).
88. Mongan, J., Moy, L. & Kahn, C. E. Jr Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).

Acknowledgements

The work presented in Fig. 5 has been funded by the Dutch Research Council (NWO) Applied and Engineering Sciences (project number 17970).

Author contributions

B.C.S. researched data for the article. All authors contributed substantially to discussion of the content. B.C.S., M.S. and A.H.M.v.d.H.-v.M. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Rheumatology* thanks Reza Forghani and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links

CS230 Deep Learning tutorial: https://cs230.stanford.edu/
List of commercial AI products for radiology: https://grand-challenge.org/aiforradiology/
Machine Learning Glossary: https://developers.google.com/machine-learning/glossary