Multilevel Modeling of Joint Damage in Rheumatoid Arthritis

Hongyang Li[1,*,+], Yuanfang Guan[1,*]

1. Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

* Contacts: hyangl@umich.edu or gyuanfan@umich.edu

+ Current affiliation: Thomas J. Watson Research Center, Yorktown Heights, NY USA

Abstract

While most deep learning approaches are developed for single images, in real world applications, images are often obtained as a series to inform decision making. Due to hardware (memory) and software (algorithm) limitations, few methods have been developed to integrate multiple images so far. In this study, we present an approach that seamlessly integrates deep learning and traditional machine learning models, to study multiple images and score joint damages in rheumatoid arthritis. This method allows the quantification of joining space narrowing to approach the clinical upper limit. Beyond predictive performance, we integrate the multilevel interconnections across joints and damage types into the machine learning model and reveal the cross-regulation map of joint damages in rheumatoid arthritis.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/aisy.202200184.

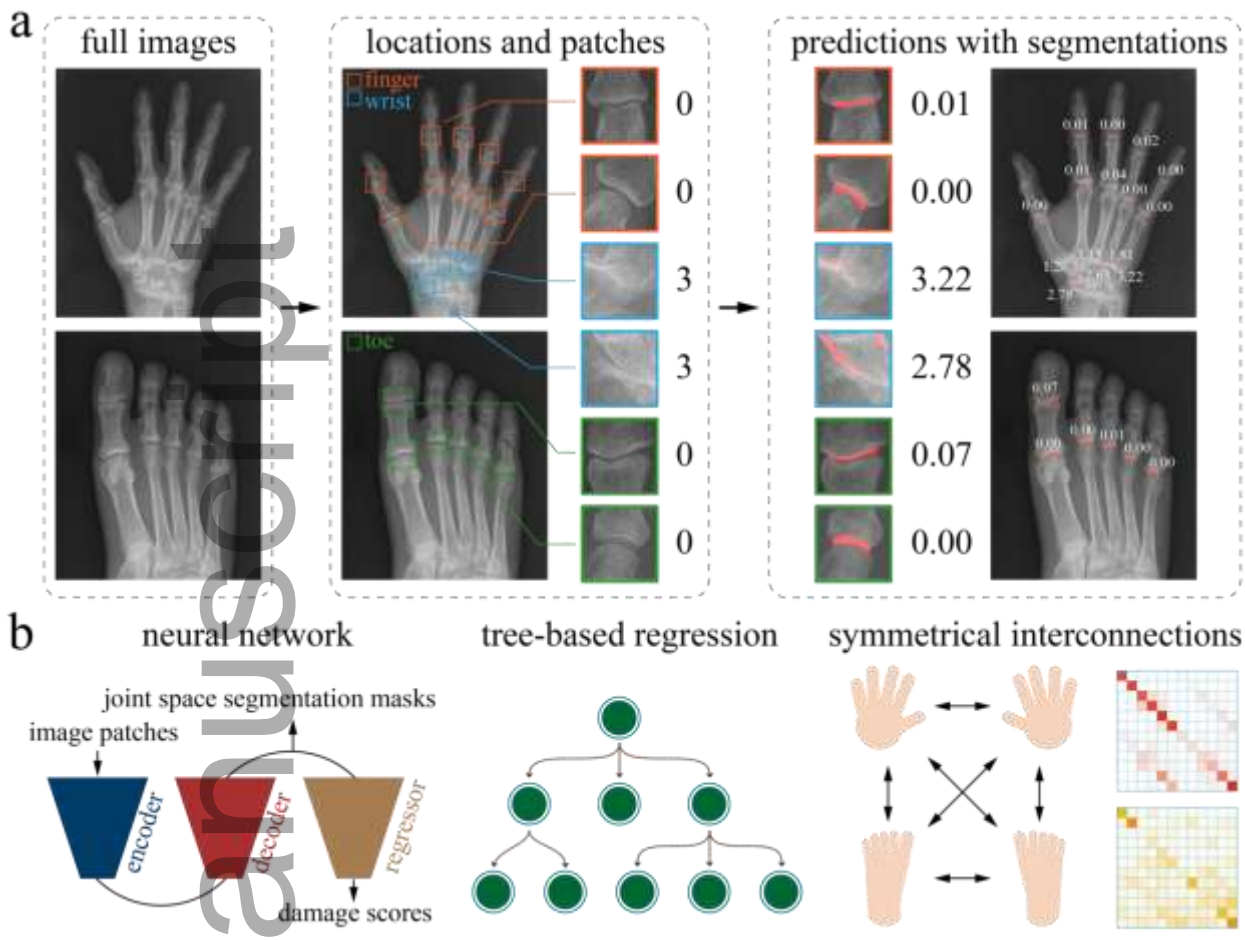

Fig. 1

Schematic representation of Mandora prediction of joint damages. a, The objective of this study is to automatically detect joint space narrowing and bone erosion damages based on radiographic images in rheumatoid arthritis. The ground truth is the SvH scores manually labeled by human scorers. Our machine learning approach, Mandora, locates joint positions and quantifies joint-specific damage levels with segmentations of joint space regions. b, We use neural network and tree-based regression models to learn contextual information of images as well as the symmetrical interconnections among joints in rheumatoid arthritis.

Introduction

Deep learning algorithms have been widely used to address biological and biomedical imaging problems in recent years [1–4]. Common image-related tasks include detection of cell nuclei [5–7], semantic segmentation of tumors [8–11], and diagnosis of diseases [12–17]. In general, standard computer vision tasks are straightforward, such as object detection, segmentation, or classification. A unique challenge posed by medical images is that there

are often multiple related images in a series, which requires clinicians to analyze and diagnose diseases through different angles across images [18–20]. For example, mammograms are often performed from both the cranial caudal view and the mediolateral oblique view [21,22]. How to integrate information from multi-view images and multiple regions is currently an active research topic in the field [23–26].

Here we focus on images of rheumatoid arthritis (RA), which is a chronic autoimmune disease that affects many regions of the body [27]. About 0.5 percent of adults are affected by RA worldwide with a higher prevalence in women than men [28]. RA primarily targets and damages joints, including pain and swelling around the joint regions. Without timely diagnosis and appropriate treatment, RA can lead to severe and irreparable joint damages and disability [27].

Many modern imaging techniques have been developed for the visualization and detection of joint damages in RA. The current standard quantification of RA damage is the Sharp/van der Heijde (SvH) scoring system with radiographic imaging [29], including joint space narrowing and bone erosion. However, manually inspecting radiographic images and evaluating SvH scores is time-consuming and effort-intensive. The fast advancement of machine learning and artificial intelligence (AI) has opened a new avenue for automatic diagnosis of RA using computers [30]. Deep learning methods have been developed to address the image-based RA scoring problems [31–34]. Yet independent and comprehensive benchmarking are needed to evaluate the predictive performance of AI methods, especially on stringently held-out testing data. Moreover, beyond high predictive performance, a key question is to gain a deeper understanding of AI methods in RA. A typical presentation of RA is symmetrical symptoms in multiple large and small joints [35,36]. Unfortunately, how to leverage the multiple images and reveal the working mechanism underlying a computational method is largely unexplored.

Here we present Mandora (**MA**chine lear**N**ing **D**etection of j**O**int damages in **R**heumatoid **A**rthritis), a machine learning approach for quantification of joint damages in RA based on radiographic images. Benefiting from cutting-edge deep learning techniques as well as conventional machine learning methods to exploit different types of information, this method automatically quantifies the degree of joint damage with high accuracy. It ranked first in scoring joint space narrowing in the recent DREAM Rheumatoid Arthritis Challenge, where state-of-the-art methods were systematically compared and benchmarked on independent

testing data. Additionally, this method segments and highlights the joint space regions to assist further disease diagnosis in clinical practice. Most importantly, we leverage the multi-level symmetrical patterns in RA patients and integrate information across multiple images to improve performance and reveal the predictive relationships across joints, damage types, and left-right sides [37,38].

Experimental Section/Methods

Data collection

In this study, we used radiographic images from two clinical studies, the Consortium for the Longitudinal Evaluation of African-Americans with Rheumatoid Arthritis (CLEAR) [39] and the Treatment Efficacy and Toxicity in Rheumatoid Arthritis Database and Repository (TETRAD) [40]. A total of 1472 radiographic images from 368 sets were used to develop models. Each set consisted of two images of hands and two images of feet from the same individual. Two types of joint damages were investigated: joint space narrowing and bone erosion. The ground truth label is the Sharp/van der Heijde (SvH) score generated by human experts through manual inspection of images [41]. Typically targeted joints by RA were examined by the SvH scoring system, including multiple joints in wrists, proximal interphalangeal (PIP), and metacarpal phalangeal (MCP) of the fingers, PIP and metatarsal phalangeal (MTP) of the toes. For joint space narrowing, 15 joints from each hand and 6 joints from each foot were assessed, with the score ranging from 0 to 5. A higher score represents a higher degree of damage, where 0 is normal, 1 is focal narrowing, 2 is the reduction of less than 50% of the original joint space, 3 is the reduction of more than 50% of the original joint space, and 4 is complete dislocation. For bone erosion, 16 joints from each hand and 6 joints from each foot were assessed, with the score ranging from 0 to 5. Similarly, the score of 0 represents no damage, 1 is discrete erosion, 2-3 are large erosions that involve the bone surface, 4 is erosion that extends over the middle of the bone, and 5 is a complete collapse. To estimate the clinical upper limit of the damage scoring problems, each joint damage was scored independently by two trained professionals. Pearson's correlation and root mean square error (RMSE) between two scorers were calculated as the clinical upper limit.

Location of joints and segmentation of joint space regions

For each joint of interest, we first manually labeled the coordinate of the center. Then based on the center coordinate, we generated a 30-by-30 pixelwise square mask as the ground truth. These location masks were used to build models to locate a joint. In addition, we also manually labeled the joint space regions using polygons. These polygons served as the segmentation masks of joint space regions, which were used to build models that accepted image patches as input.

Evaluation and experimental design

Two evaluation metrics were considered to assess the performance of predicting joint damages. The first metric is Pearson's correlation coefficient (r) defined as follows:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \underline{x})(y_i - \underline{y})}{S_x S_y}$$

where n is the number of joint damages to be scored, x_i is the prediction, y_i is the ground truth label created by the human scorer, $\underline{x}$ and $\underline{y}$ are the averages, S_x and S_y are the standard deviations. The secondary metric is RMSE defined as follows:

$$RMSE = \sum_{i=1}^{n}(x_i - y_i)^2/n$$

where n is the number of joint damages to be scored, x_i is the prediction, and y_i is the ground truth label. To evaluate the performance of a model, we performed 10-fold cross-validation experiments, where 90 percent of the data were used to build models and 10 percent of the data were held out for testing.

In addition to predicting joint damage, we also evaluated the accuracy of locating a joint. Since image sizes varied across individuals, we normalized the distance measurement during evaluation. Specifically, the coordinate of a point was first scaled to a continuous value between 0 and 1 by being divided by the height and width of an image. Then we

calculated the distance between predictions and ground truth labels. Using this normalized distance instead of pixel values, images with different sizes were comparable.

Deep learning prediction of joint locations based on full images

We designed a deep convolutional neural network for predicting joint locations. This neural network contains an encoder and a decoder with multiple convolutional layers, max-pooling layers, up-convolutional layers, and concatenation layers. The kernel sizes of convolutional layers, max-pooling layers, and up-convolutional layers are 7×7, 2×2, 2×2, respectively. The encoder has 8 convolutional layers as well as 4 max-pooling layers. The decoder has 8 convolutional layers and 4 up-convolutional layers. Meanwhile, the encoder and the decoder are connected through concatenation layers. Each convolutional layer is followed by a non-linear "ReLU" activation, except for the last convolutional layer that is activated by a "sigmoid" function. The batch normalization layer is applied before each convolutional layer to accelerate the training process. The cross-entropy loss was used together with the Adam optimizer. The neural network was first trained 50 epochs with the learning rate of 0.001, then trained 50 epochs with a smaller learning rate of 0.0001. The neural network was implemented using "Keras" (2.2.4) with the "Tensorflow" (1.14.0) backend in Python.

Deep learning prediction of joint damages based on image patches

We designed a special deep convolutional neural network for predicting joint damages based on image patches. Specifically, this neural network has a encoder-decoder-regressor architecture with two outputs: (1) segmentation of the joint space region, and (2) joint damage score. The encoder has 8 convolutional layers and 4 max-pooling layers. The decoder has 8 convolutional layers and 4 up-convolutional layers. The regressor has 8 convolutional layers and 4 max-pooling layers. Meanwhile, concatenation layers were used to link (1) the encoder and the decoder, and (2) the decoder and the regressor. The decoder outputs the joint space segmentation that is used as the input for the regressor as well. The regressor outputs the joint damage score. The kernel sizes of convolutional layers, up-convolutional layers, and max-pooling layers are 7×7, 2×2, 2×2, respectively. Each convolutional layer is followed by a non-linear "ReLU" activation. The batch normalization layer is applied before each convolutional layer to accelerate the training process. The last layer is a "Dense" layer that is activated by a "sigmoid" function. The cross-entropy loss was

used together with the Adam optimizer. The neural network was first trained 10 epochs with the learning rate of 0.001, then trained 40 epochs with a smaller learning rate of 0.0001. The neural network was implemented using “Keras” (2.2.4) with the “Tensorflow” (1.14.0) backend in Python.

Tree-based learning of symmetrical patterns

To learn the symmetrical patterns among joints and damage types, we further build a tree-based machine learning model. Specifically, for each joint damage to be predicted, the image patch-based predictions of all joint damages from the neural network model are used as the input. The tree-based model automatically learns the non-linear relationships among all joints from hands and feet, and two types of joint damages. A total of 500 trees are used in the ensemble predictions with a maximum depth of 4. The tree-based model was implemented using the “etr” function of “scikit-learn” (0.21.2) in Python.

SHAP analysis

To investigate the interconnections among joints and damage types in our method, we performed the SHAP analysis of the tree-based model. For each joint damage of interest, the absolute SHAP values of all joint damages are calculated, representing the strength of contribution to determining the degree of damage. The joint-wise absolute SHAP values are shown as two heatmaps: (1) the contributions of joints from the same side, and (2) the contributions from the other side. The SHAP analysis was performed using “shap” (0.31.0) in Python.

Statistical analysis

To test whether two results are significantly different in cross-validation experiments, we performed the one-sided paired Wilcoxon signed-rank test using R (3.6.1).

Ethical approval

Ethical approval was done through Sage Bionetworks who distributed the data through the cloud environment.

Data availability

The radiographic images used in this study were downloaded from the challenge website:

https://www.synapse.org/#!Synapse:syn20545111/wiki/597243

The segmentation masks of the joint space regions are available on our website:

https://guanfiles.dcmb.med.umich.edu/Mandora/segmentation_joint_space

Code availability

The code of Mandora is available in the GitHub repository:

https://github.com/GuanLab/Mandora

Author contributions

HL and YG conceived and designed this project and wrote the paper. HL implemented the method, performed the experiments

Results

Overview of experimental design

Scoring joint damages in RA is a complex task with two subtasks: (1) object detection - we need to detect and locate multiple joints within an image, and (2) disease recognition - we need to predict the degree of damage through regression analysis. To solve this unique problem, we developed a multi-step pipeline (Fig. 1). We first built a deep convolutional neural network model to identify the location of each joint. Once we obtained the location, an image was cut into small patches that were centered around joints. Then the image patches were used as the input for a specially designed neural network to perform regression analysis of joint damage scores, as well as semantic segmentation of joint space regions. Notably, segmentation was not required, yet it significantly improved predictive performance. Finally, patients with RA are likely to develop symmetrical symptoms in both sides of hands and feet, and joint space narrowing and bone erosion often go hand in hand. We therefore developed a tree-based conventional machine learning model to integrate all available information from both sides and two types of damages. The step not only further improved the performance but also revealed cross-joint prediction relationships that are the nature of RA.

Convolutional neural network locates joint positions with high accuracy

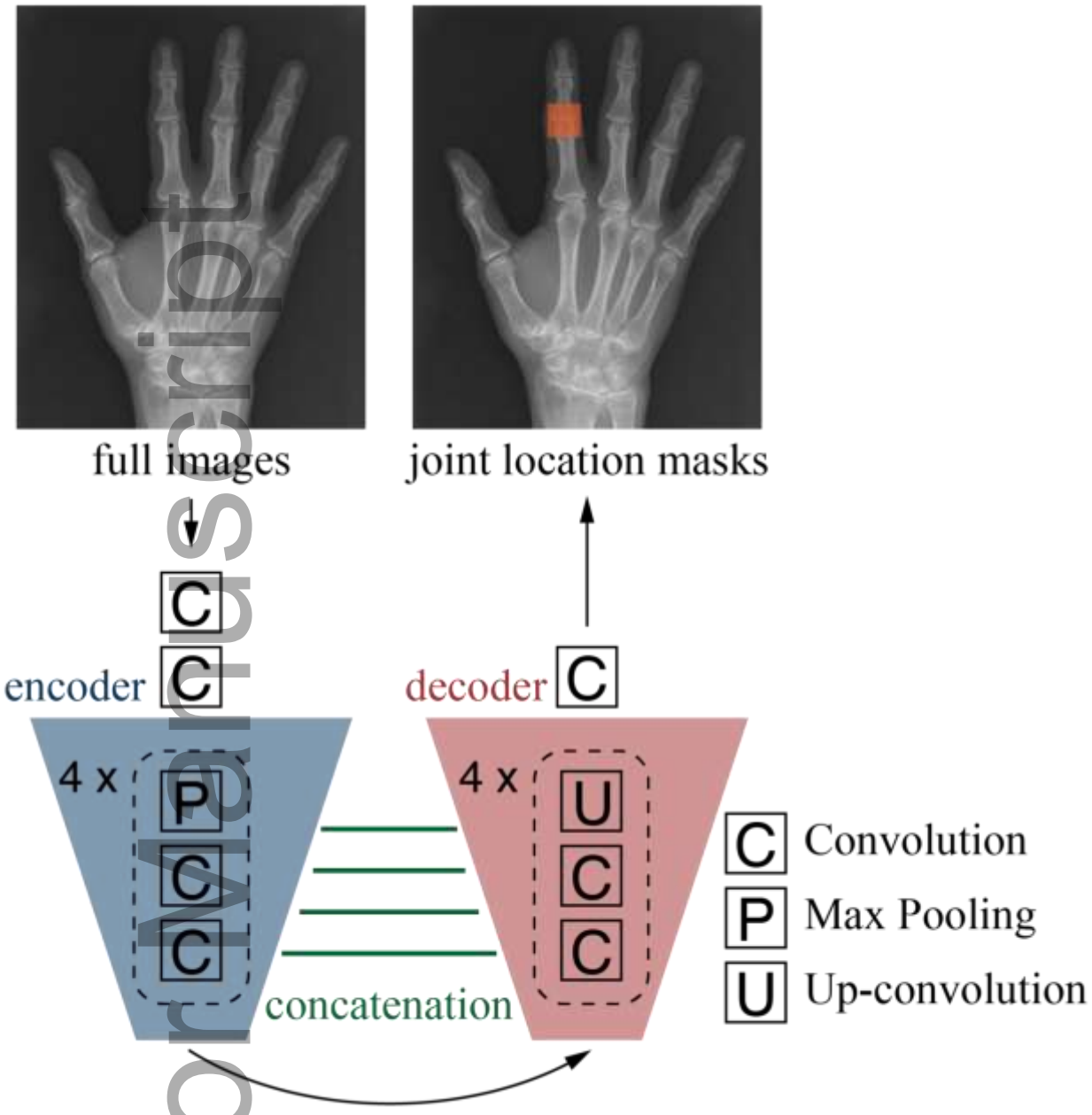

Fig. 2

Architecture of the deep convolutional neural network in localizing joint positions. This neural network contains an encoder to extract information at multiple scales and a decoder to decode the abstracted information from feature maps. Multiple convolution, max-pooling and up-convolution layers are used. The encoder and the decoder are connected through concatenation layers.

We first built a 2D convolutional neural network model to detect the location of a joint (Fig. 2). Specifically, for each joint to be located, we used a 30-by-30 pixel-wise square mask as the ground truth label to present the location and the entire image as the input. This

convolutional neural network has an encoder that extracts feature maps at multiple resolutions and scales through convolutional layers, as well as a decoder that decrypts the abstract information within feature maps through up-convolutional layers. Meanwhile, concatenation layers are used to link the encoder and decoder to prevent information decay.

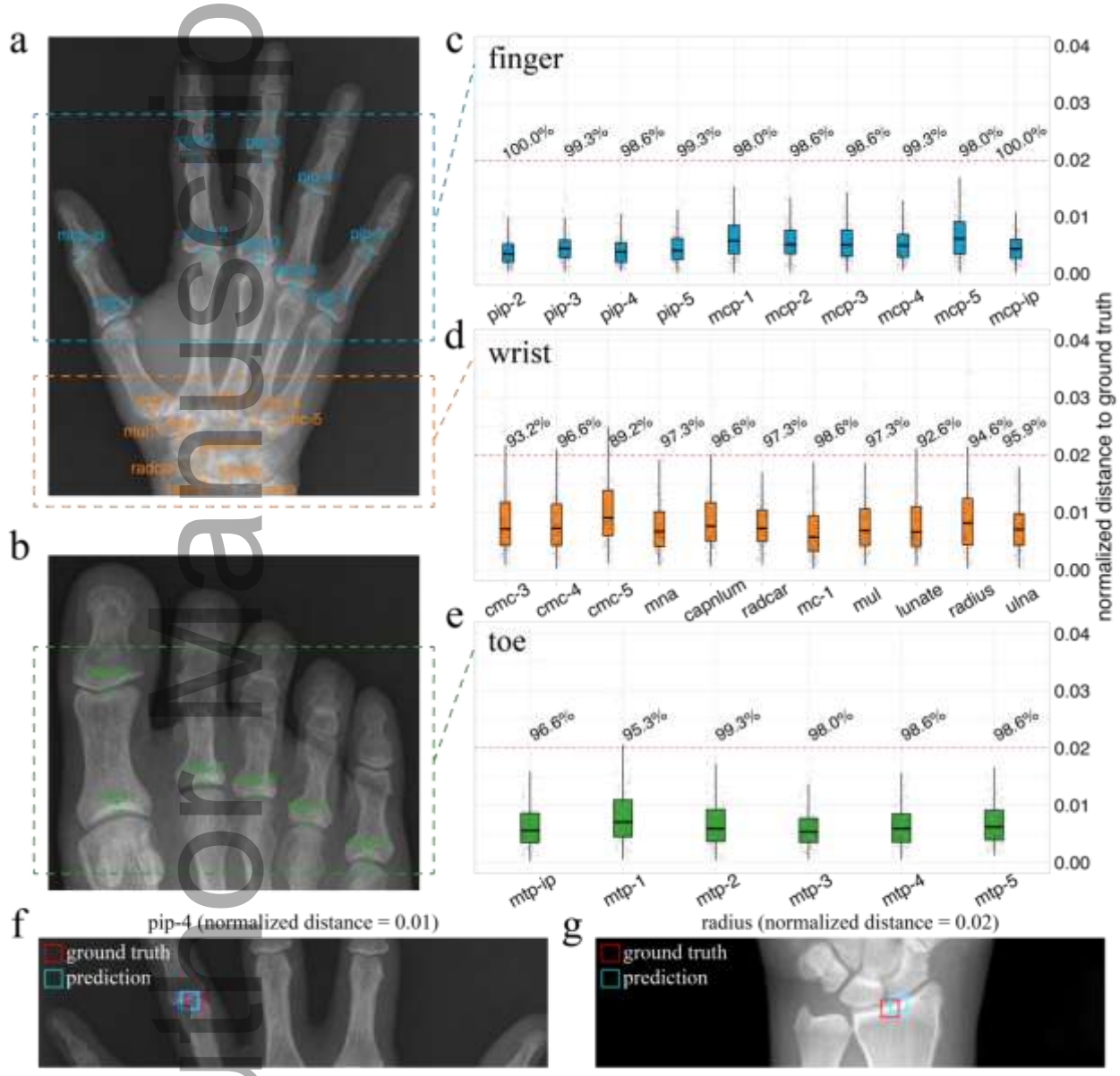

Fig. 3

Mandora localizes joint positions with high accuracy. a-b, We develop a deep convolutional neural network model for detecting locations of multiple joints from finger, wrist, and toe. Instead of using the pixel as the unit, we measure the difference between prediction and ground truth using a normalized distance so that images with different sizes are comparable. c-e, The predictive performances of locating joints from the finger, wrist, and toe are shown as box plots. The horizontal dashed red line represents the cutoff, a normalized distance of

0.02. The numbers represent the percentage of samples with a smaller distance than the cutoff. f-g, Two examples are shown with normalized distances of 0.01 and 0.02, respectively.

To evaluate the predictive performance of joint location, we performed 10-fold cross-validation experiments for each joint in the finger, wrist (Fig. 3a), and toe (Fig. 3b). Since the sizes vary across images, we used a normalized distance to measure the difference between predictions and ground truth labels (see details in Methods). Briefly, the coordinates of each point within an image are rescaled from pixel values to a continuous value between 0 and 1 by dividing the height or width of an image, so that the results are uniform and comparable across images with different sizes. The distributions of normalized distances are shown as boxplots in Fig. 3c-e. We selected a normalized distance of 0.02 as the cutoff (horizontal red dashed lines) to measure the predictive accuracy. Examples of normalized distances of 0.01 and 0.02 are shown in Fig. 3f-g. In general, joints from fingers are easier to locate and more than 98.0% of testing joints are within the 0.02 normalized distance (Fig. 3c). In contrast, joints from wrists are harder to distinguish, owing to their proximity (Fig. 3d). The accuracy of locating joints from toes is similar to that from fingers (Fig. 3e). Overall, the convolutional neural network model locates joints with high accuracy.

Segmentation of joint space regions significantly improves joint damage prediction

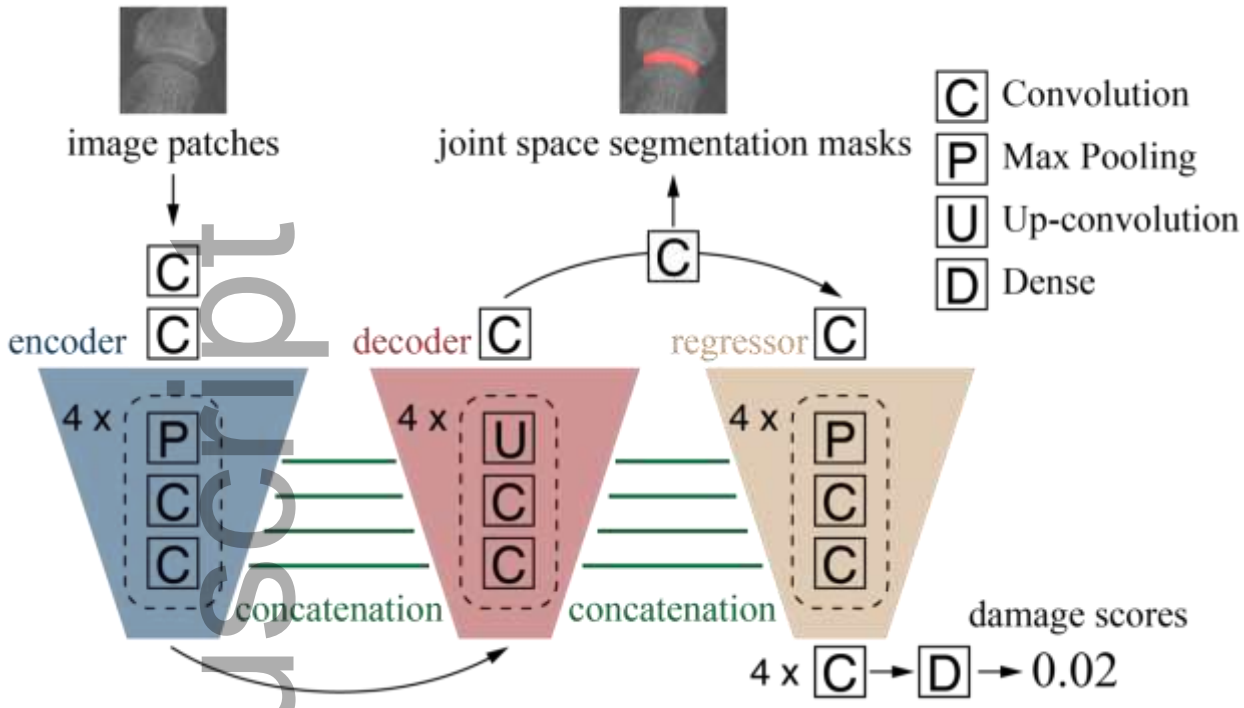

Fig. 4

Architecture of the deep neural network for segmenting joint space regions and predicting damage scores. This neural network contains an encoder to extract information at multiple scales, a decoder to decode the abstracted information from feature maps, and a regressor to quantify the damage score of a joint. Multiple convolution, max-pooling, up-convolution layers and one dense layer are used. The decoder generates the segmentation mask, which is further used as input for the regressor. The encoder, the decoder and the regressor are connected through concatenation layers.

Based on the joint location obtained from the previous step, we cut full images into image patches that were centered around joints to be scored. Then the patches were treated as the input for a deep learning model to predict the damage score of each joint. We design a novel neural network architecture for the patch-based damage prediction that simultaneously outputs the damage score as well as the segmentation of the joint space region (Fig. 4). Specifically, the architecture contains two parts. The first part includes an encoder and a decoder so that it extracts features from multiple scales and resolutions. The output of the first part is the segmentation mask that is further used as the input for the second part of the neural network. The rationale is that with the guidance of the segmentation mask, the neural network will easily learn where to “look at” and focus primarily on the regions of interest to determine the damage level. The second part contains a regressor that generates one output value representing the damage score.

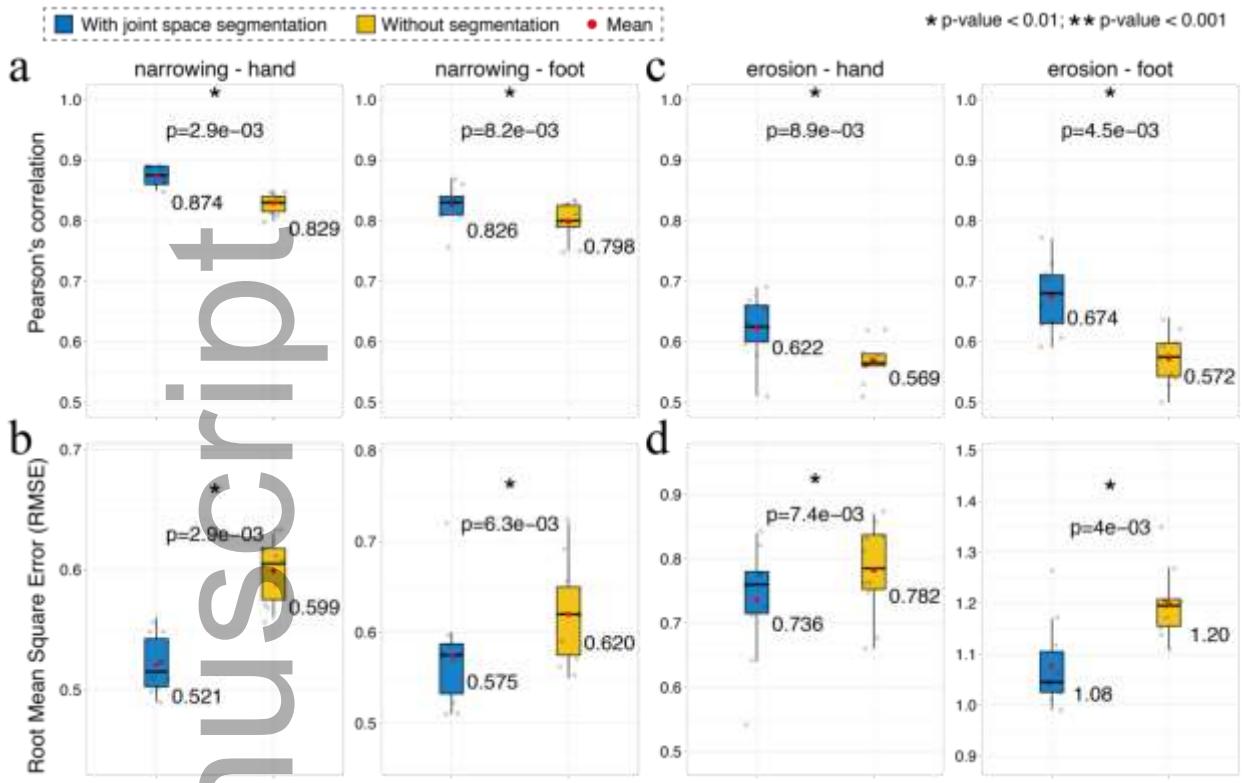

Fig. 5

Improving the predictive performance with semantic segmentation of joint space regions. We compared two types of neural network models with or without the semantic segmentation of joint space regions. We benchmarked their performance in predicting joint space narrowing using a, Pearson's correlation and b, root mean square error. We also benchmarked their performance in bone erosion prediction using c, Pearson's correlation and d, root mean square error. In each comparison, we performed 10-fold cross-validation experiments. The one-sided paired Wilcoxon signed-rank tests were used to determine the statistical significance.

To comprehensively evaluate the predictive performance and investigate the effect of the special neural network architecture, we performed 10-fold cross-validation experiments on hands and feet individually. The primal evaluation metric is Pearson's correlation between predictions and ground truth labels created by human professionals. We also considered a secondary metric, the root mean square error (RMSE) between prediction and ground truth. We benchmarked the neural network models with or without the segmentation of the joint space region. In both hands and feet, the neural network with segmentation achieved significantly higher correlations than the network without segmentation, as well as lower RMSEs (Fig. 5a-b).

Since joint space narrowing and erosion often occur simultaneously, a joint with narrowing damage is likely to have bone erosion. We therefore hypothesize that the neural network architecture with segmentation will also improve the prediction accuracy of erosion damages. Similar to joint space narrowing, this model significantly increased the correlations and decreased RMSEs for erosion prediction in both hands and feet (Fig. 5c-d).

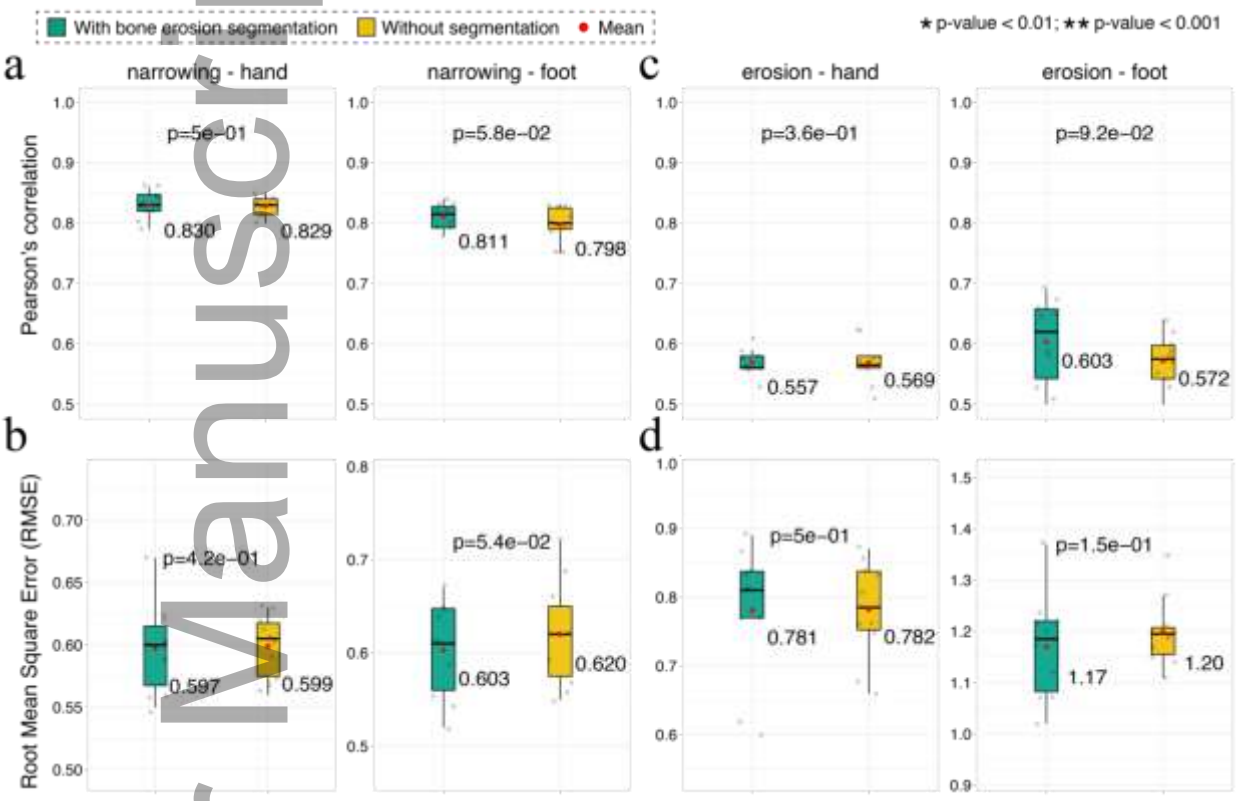

Fig. 6

Benchmarking predictive performance of neural network models with or without segmentations of bone erosion. We compared two types of neural network models with or without the semantic segmentation of joint space regions. We benchmarked their performance in predicting joint space narrowing using a, Pearson's correlation and b, root mean square error. We also benchmarked their performance in bone erosion prediction using c, Pearson's correlation and d, root mean square error. In each comparison, we performed 10-fold cross-validation experiments. The paired Wilcoxon signed-rank tests were used to determine the statistical significance.

Intuitively, integrating segmentation of bone erosion regions should also improve performance similar to including segmentation of joint space regions. Unfortunately, the

segmentation of erosion did not improve the performance (Fig. 6). The main reason is that the segmentation of erosion is much harder than narrowing and the quality of segmentation is relatively low. We anticipate that this neural network will further improve with high-quality erosion masks in future studies.

Mandora reveals joint-wise symmetrical contributions to damage prediction

The observed correlations of damage types and joint locations across individuals.

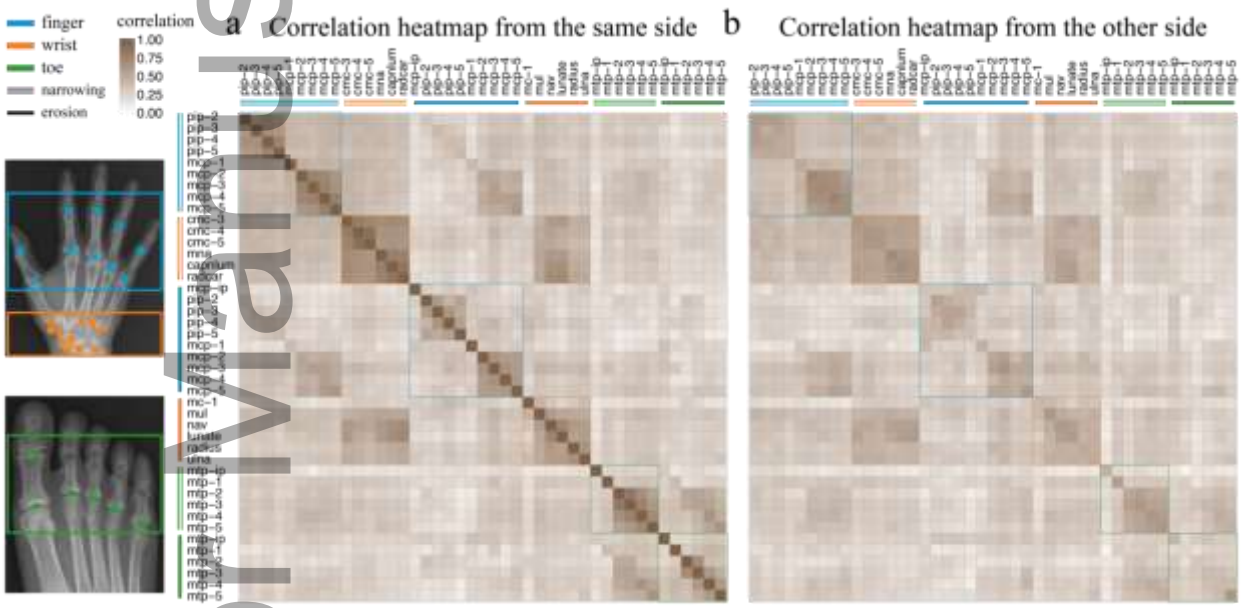

Fig. 7

Correlation analysis of joint damages in rheumatoid arthritis. For each joint damage, we calculated the pairwise Pearson's correlation from all other joint damages. Joints from finger, wrist, toe are shown in blue, orange, and green, respectively. The double lines represent joint space narrowing and the single line represents bone erosion. a, The correlation heatmap represents the correlations from the same side. b, The correlation heatmap represents the correlations from the other side.

Patients with RA often develop multi-level symmetrical and correlated joint damages in both hands and feet. In fact, through analyzing total joint damage scores at the image level, we observed medium to high Pearson's correlations at three aspects: (1) between joint space

narrowing and bone erosion, (2) between the left side and right side, and (3) between hands and feet (Table 1). We also calculated the cross-patient pairwise correlations among all joints and found ubiquitous correlations shown as heatmaps in Fig. 7.

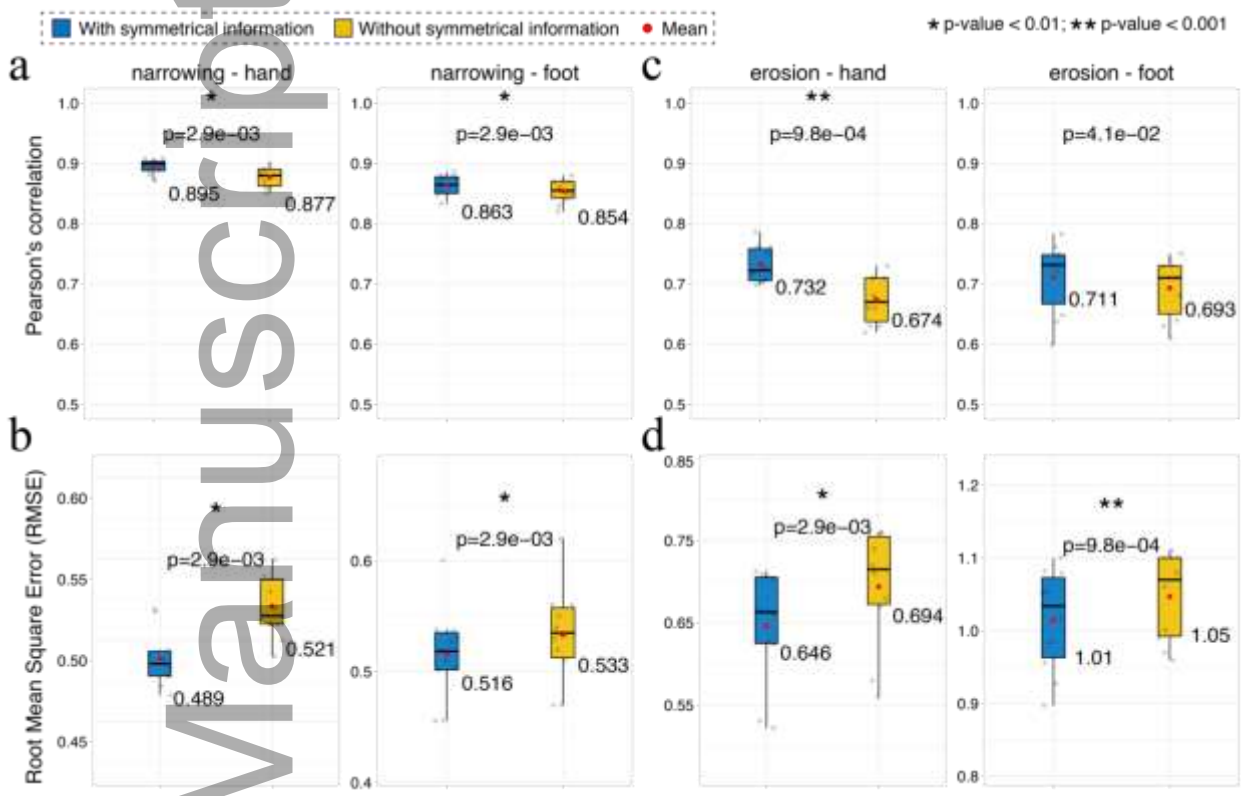

Fig. 8

Improving the predictive performance with symmetrical information. We compared machine learning approaches with or without integrating the symmetrical information. We benchmarked their performance in predicting joint space narrowing using a, Pearson's correlation and b, root mean square error. We also benchmarked their performance in bone erosion prediction using c, Pearson's correlation and d, root mean square error. In each comparison, we performed 10-fold cross-validation experiments. The one-sided paired Wilcoxon signed-rank tests were used to determine the statistical significance.

To leverage this unique characteristic feature and consider the multi-level associations in RA, we developed a conventional tree-based machine learning model. Specifically, for each joint, the image patch-based predictions of all joint damages from the previous step are treated as input features in the tree-based model. The associations and dependent relationships among joints are learned automatically in this machine learning model. It

further significantly improved the predictive performance of joint space narrowing and bone erosion in both hand and foot, except for the Pearson's correlation of erosion in the foot (Fig. 8). This model indeed improved the correlation from 0.693 to 0.711, yet the difference is not statistically different based on the one-sided paired Wilcoxon signed-rank test (p-value = 0.041).

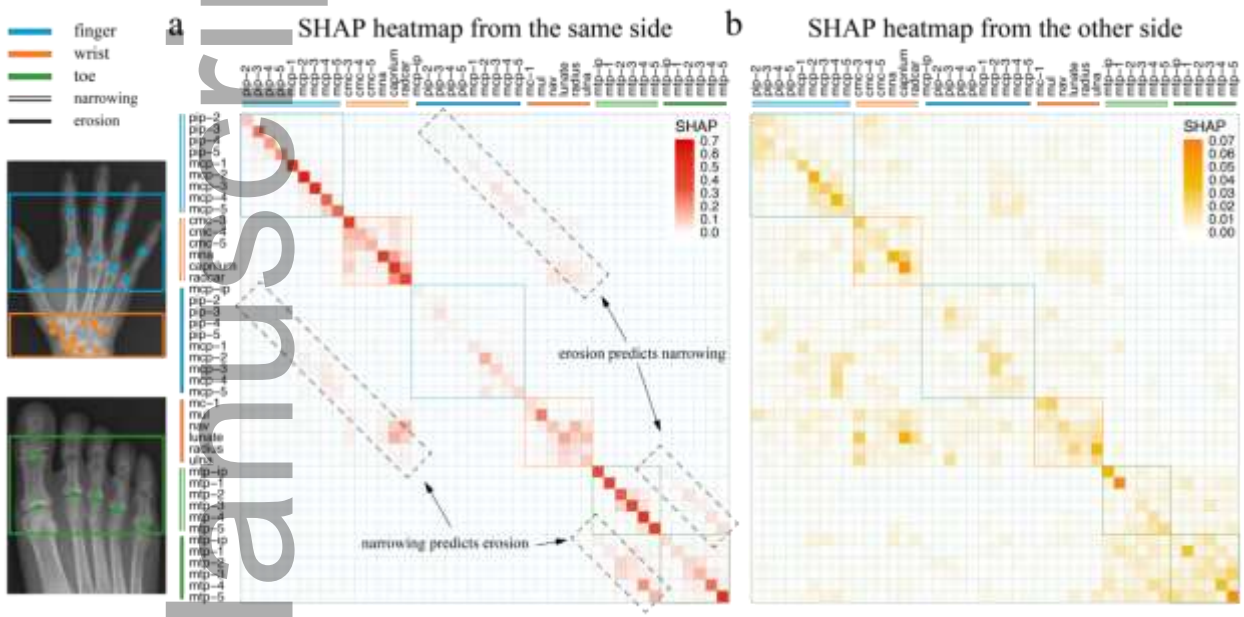

Fig. 9

SHAP analysis reveals the interconnection heatmaps among joint damages in Mandora. We performed SHAP analysis of the tree-based models in Mandora. For each joint damage, we calculated the absolute SHAP values from all joints. Joints from the finger, wrist, and toe are shown in blue, orange, and green, respectively. The double lines represent joint space narrowing and the single line represents bone erosion. The joints along the row are the target being predicted, whereas joints along the column serve as features to predict. a, The SHAP heatmap represents the contributions from the same side. b, The SHAP heatmap represents the contributions from the other side.

Beyond high performance, an essential topic of machine learning studies is to dissect the "black box" method and understand how it works. To reveal the regulatory relationships among joints in our model, we further performed SHAP analysis and presented the results as a heatmap of the average absolute SHAP values (Fig. 9a and Table 2). In this heat map, the joints in the column serve as features to predict other joints, whereas the joints in the

row are the targets being predicted. The color bars represent their locations in the finger, wrist, or toe. Double lines represent joint space narrowing and a single line represents bone erosion. Higher SHAP values demonstrate more important contributions in predicting the damage scores. As expected, the strongest predictor is generally the joint itself, which is shown as the red diagonal trace in Fig. 9a. Joints from the wrist are often predictable by each other due to their proximity, shown as clustered redness in two submatrices (narrowing and erosion) with orange borders. Intriguingly, we also observed off-diagonal red traces highlighted in four dashed rectangles. For example, the very top dashed rectangle displays the contribution of bone erosion to predict joint narrowing in hand - if a joint has erosion, it is more likely to have narrowing. Altogether, these four rectangles manifest the interconnections between narrowing and erosion in both hands and feet. In addition to the regulation among joints from the same side, we also investigated the contributions of joints from the other side. Again, we observed a stronger diagonal trace, indicating the special and important role of the counterpart joint from the other side in predicting joint damage (Fig. 9b and Table 3). In summary, the SHAP analysis reveals the working mechanisms underlying our machine learning model, especially the multidimensional associations across joint damage types and locations. Consistent with correlation analysis (Fig. 7) and the distinct pattern of joint damage distribution in RA, our method grasps useful information and empowers predictions of joint damage.

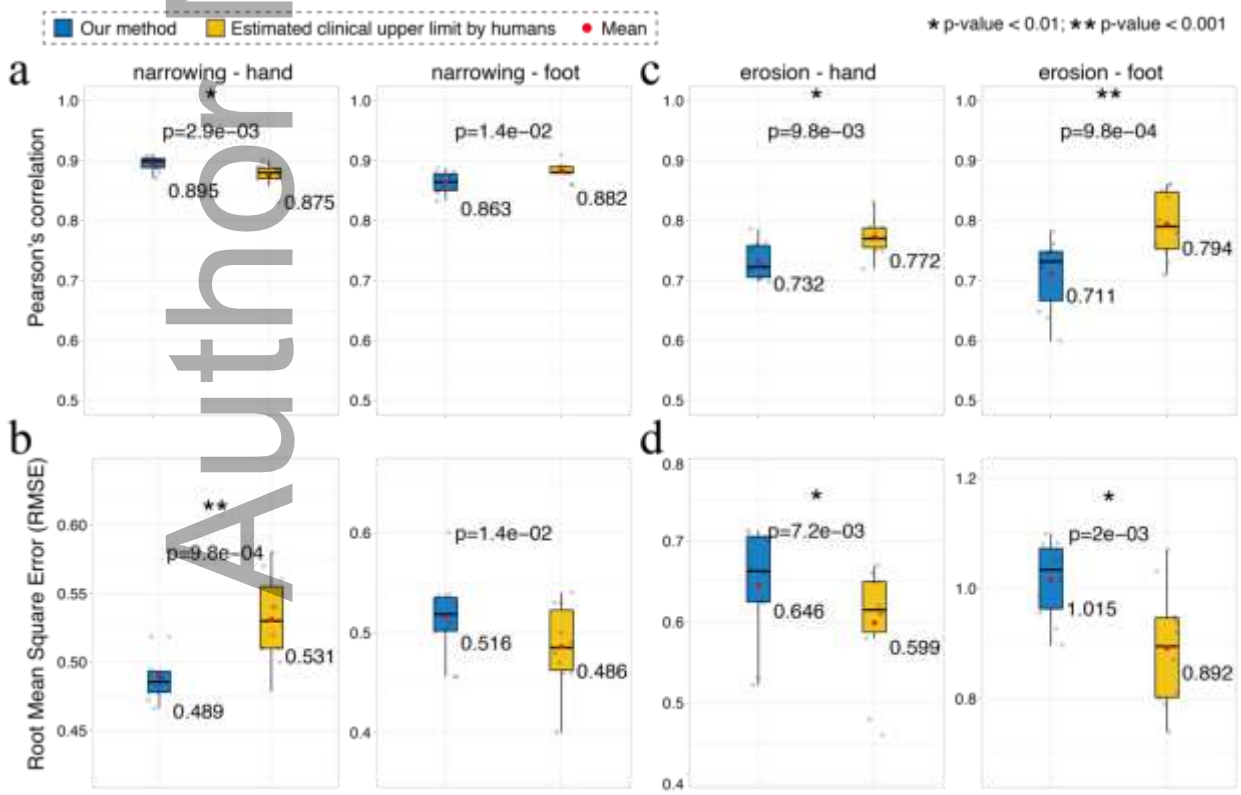

Fig. 10

Mandora predictions approach the clinical upper limit estimated by human scorers. The joint damages were scored independently by two human scorers. The correlation and RMSE between them were calculated to estimate the clinical upper limit. We benchmarked Mandora's predictions in predicting joint space narrowing using a, Pearson's correlation and b, root mean square error. We also benchmarked the performance in bone erosion prediction using c, Pearson's correlation and d, root mean square error. In each comparison, we performed 10-fold cross-validation experiments. The paired Wilcoxon signed-rank tests were used to determine the statistical significance.

Another crucial question in machine learning study is how close the performance difference is between a machine learning model and a human. In practice, human scorers are not perfect, especially for the tedious and time-consuming SvH scoring. The clinical upper limit of this RA scoring problem can be estimated by calculating the difference between two human scorers. We therefore calculated Pearson's correlation between scores from two trained human professionals. The performance of our method is compared with this clinical upper limit (Fig. 10). For joint space narrowing, our method achieved comparable performance with human scorers. For joint erosion, there is still space to improve machine learning methods. Overall, these results indicate that our method is closing the gap between computers and humans for the joint damage scoring problem, especially for joint space narrowing.

Discussion

In this study, we develop a novel machine learning approach for integrating multiple images and automatically quantifying joint space narrowing and bone erosion in rheumatoid arthritis. We designed a special neural network architecture that simultaneously scores joint damage levels and segments the joint space regions. This design not only significantly improves the prediction accuracy but also highlights the regions of interest to assist further analysis in clinical settings. The idea of introducing segmentation into an image-based regression deep learning model should not be limited to the joint damage scoring in RA. In fact, many biomedical imaging problems have a similar situation - the quantification or diagnosis closely depends on parts of an image. The segmentation of the disease-related regions of interest will be crucial to guide a neural network to focus on those regions, improve the performance, and facilitate subsequent error analysis or clinical diagnosis. This is especially true when

the sample size is relatively small and the segmentation will serve as a “teacher” that helps the model training well with a limited number of samples.

Inspired by the widely observed symmetrical symptoms in patients with RA, the method we developed learns multilevel symmetry and dependence across images. This approach is novel in that it seamlessly integrates multiple layers of information from different images to guide prediction, which can be extended to other medical image fields. Additionally, we investigated the relationships among joints and damage types in our machine learning model and revealed the disease-specific map; this data-driven RA-specific map is instructive to clinical decisions. This study design can be applied to many biomedical imaging problems and biological studies, with or without symmetrical patterns. Through analyzing the contributions of different, multiple images used in a machine learning model, the hidden relationships between different disease manifestations will be revealed from a new computational perspective, complementing direct experimental observations and current knowledge.

Although many deep learning models have been developed for image-based joint damage detection, there is still room for improvement. Top-performing methods in the DREAM Rheumatoid Arthritis Challenge, including ours, consist of multiple steps [41]. Multi-step methods require more human designing of multiple modules, whereas end-to-end methods have more simplified workflows and are easier to deploy without external priors and constraints in clinical settings [42]. Ideally, end-to-end deep learning algorithms should be designed to simultaneously output damages scores of joint space narrowing as well as bone erosion. Yet the performance of end-to-end approaches without hand-engineered components will be largely limited by the much smaller size of images, compared to millions of images in ImageNet [43]. Moreover, unlike simple objective detection tasks in computer vision, the nature of detecting multiple joints and two types of damages within an image largely complicates the problem. In practice, a fully automatic deep learning system for biomedical image analysis can significantly benefit many clinical disciplines in terms of efficiency and cost-effectiveness [44]. However, complete automation remains a translational gap, where human-in-the-loop computing can be beneficial for many biomedical image problems [45].

In terms of detecting bone erosion in RA, we observe a gap between our method and the clinical upper limit. The major limitation is the relatively low quality of erosion scores. In fact, Pearson's correlation between erosion damage scores (~0.78) by two trained human

professionals is significantly lower than that of joint narrowing scores (~0.88). This indicates that scoring bone erosion damage is in nature more difficult. Therefore, more inconsistency is observed between human experts. In our computational pipeline, a key component to improve performance is the segmentation of damaged joint regions. In contrast to the straightforward segmentation of joint space regions, the determination of bone erosion regions may vary between humans. If high-quality segmentation of bone erosion as well as more consistent erosion scores are available in future studies, the prediction performance will further improve to close the gap between artificial intelligence methods and the clinical upper limit.

Conclusion

We develop an AI approach for automatic scoring joint space narrowing and bone erosion. Through semantic segmentation of the joint space region as well as integration of multilevel interconnection across joints and damage types, our method achieved high prediction accuracy in quantifying joint space narrowing, approaching the clinical upper limit of this problem.

Acknowledgements

This work is supported by NIH/NIGMS R35GM133346 and NSF/DBI #1452656.

Conflict of interest

The authors declare no competing interests.

References

[1]Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.

[2]D. Shen, G. Wu, H.-I. Suk, *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221.

[3]C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, C.-W. Lin, *Neural Netw.* **2020**, *131*, 251.

[4]J. Ker, L. Wang, J. Rao, T. Lim, *IEEE Access* **2018**, *6*, 9375.

[5]K. Sirinukunwattana, S. E. Ahmed Raza, Yee-Wah Tsang, D. R. J. Snead, I. A. Cree, N. M. Rajpoot, *IEEE Trans. Med. Imaging* **2016**, *35*, 1196.

[6]L. Yang, R. P. Ghosh, J. M. Franklin, S. Chen, C. You, R. R. Narayan, M. L. Melcher, J. T. Liphardt, *PLoS Comput. Biol.* **2020**, *16*, e1008193.

[7]D. A. Van Valen, T. Kudo, K. M. Lane, D. N. Macklin, N. T. Quach, M. M. DeFelice, I. Maayan, Y. Tanouchi, E. A. Ashley, M. W. Covert, *PLoS Comput. Biol.* **2016**, *12*, e1005177.

[8]W. Gómez-Flores, W. Coelho de Albuquerque Pereira, *Comput. Biol. Med.* **2020**, *126*, 104036.

[9]S. Wu, H. Li, D. Quang, Y. Guan, *Radiol Artif Intell* **2020**, *2*, e190011.

[10] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, Y. Fan, *Med. Image Anal.* **2018**, *43*, 98.

[11] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, *Med. Image Anal.* **2017**, *35*, 18.

[12] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, O. Ronneberger, *Nat. Med.* **2018**, *24*, 1342.

[13] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, S. Ji, *Med. Image Comput. Comput. Assist. Interv.* **2014**, *17*, 305.

[14] Y. Q. Jiang, J. H. Xiong, H. Y. Li, X. H. Yang, W. T. Yu, M. Gao, X. Zhao, Y. P. Ma, W. Zhang, Y. F. Guan, H. Gu, J. F. Sun, *Br. J. Dermatol.* **2020**, *182*, 754.

[15] Y. Guan, H. Li, D. Yi, D. Zhang, C. Yin, K. Li, P. Zhang, *Nat Comput Sci* **2021**, *1*, 433.

[16] H. Li, Y. Guan, *Commun Biol* **2021**, *4*, 18.

[17] H. Zhang, K. Deng, H. Li, R. L. Albin, Y. Guan, *Patterns (N Y)* **2020**, *1*, DOI 10.1016/j.patter.2020.100042.

[18] D. Rey, G. Subsol, H. Delingette, N. Ayache, *Med. Image Anal.* **2002**, *6*, 163.

[19] M. A. Mazurowski, M. Buda, A. Saha, M. R. Bashir, *J. Magn. Reson. Imaging* **2019**, *49*, 939.
[20] R. J. Radke, S. Andra, O. Al-Kofahi, B. Roysam, *IEEE Trans. Image Process.* **2005**, *14*, 294.
[21] A. M. J. Bluekens, R. Holland, N. Karssemeijer, M. J. M. Broeders, G. J. den Heeten, *Radiology* **2012**, *265*, 707.
[22] S. Zackrisson, K. Lång, A. Rosso, K. Johnson, M. Dustler, D. Förnvik, H. Förnvik, H. Sartor, P. Timberg, A. Tingberg, I. Andersson, *Lancet Oncol.* **2018**, *19*, 1493.
[23] Y. Wu, X. Han, Y. Su, M. Glidewell, J. S. Daniels, J. Liu, T. Sengupta, I. Rey-Suarez, R. Fischer, A. Patel, C. Combs, J. Sun, X. Wu, R. Christensen, C. Smith, L. Bao, Y. Sun, L. H. Duncan, J. Chen, Y. Pommier, Y.-B. Shi, E. Murphy, S. Roy, A. Upadhyaya, D. Colón-Ramos, P. La Riviere, H. Shroff, *Nature* **2021**, *600*, 279.
[24] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, T. Funkhouser, **2021**, DOI 10.48550/ARXIV.2102.13090.
[25] G. Carneiro, J. Nascimento, A. P. Bradley, in *Deep Learning for Medical Image Analysis*, Elsevier, **2017**, pp. 321–339.
[26] Y. Guan, X. Wang, H. Li, Z. Zhang, X. Chen, O. Siddiqui, S. Nehring, X. Huang, *Patterns (N Y)* **2020**, *1*, DOI 10.1016/j.patter.2020.100106.
[27] J. S. Smolen, D. Aletaha, A. Barton, G. R. Burmester, P. Emery, G. S. Firestein, A. Kavanaugh, I. B. McInnes, D. H. Solomon, V. Strand, K. Yamamoto, *Nat Rev Dis Primers* **2018**, *4*, 18001.
[28] D. Aletaha, J. S. Smolen, *JAMA* **2018**, *320*, 1360.
[29] D. M. van der Heijde, M. A. van Leeuwen, P. L. van Riel, L. B. van de Putte, *J. Rheumatol.* **1995**, *22*, 1792.
[30] K. M. Kingsmore, C. E. Puglisi, A. C. Grammer, P. E. Lipsky, *Nat. Rev. Rheumatol.* **2021**, *17*, 710.
[31] B. Stoel, *RMD Open* **2020**, *6*, DOI 10.1136/rmdopen-2019-001063.
[32] Y. Huo, K. L. Vincken, D. van der Heijde, M. J. H. De Hair, F. P. Lafeber, M. A. Viergever, *IEEE Trans. Biomed. Eng.* **2016**, *63*, 2177.
[33] S. Murakami, K. Hatano, J. Tan, H. Kim, T. Aoki, *Multimed. Tools Appl.* **2018**, *77*, 10921.
[34] J. Rohrbach, T. Reinhard, B. Sick, O. Dürr, *Comput. Electr. Eng.* **2019**, *78*, 472.
[35] Q. Guo, Y. Wang, D. Xu, J. Nossent, N. J. Pavlos, J. Xu, *Bone Res* **2018**, *6*, 15.
[36] V. Majithia, S. A. Geraci, *Am. J. Med.* **2007**, *120*, 936.
[37] L. Shapley, *Classics in Game Theory* **1997**, 69.

[38] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, S.-I. Lee, *Nat Biomed Eng* **2018**, *2*, 749.
[39] S. L. Bridges Jr, Z. L. Causey, P. I. Burgos, B. Q. N. Huynh, L. B. Hughes, M. I. Danila, A. van Everdingen, S. Ledbetter, D. L. Conn, A. Tamhane, A. O. Westfall, B. L. Jonas, L. F. Callahan, E. A. Smith, R. Brasington, L. W. Moreland, G. S. Alarcón, D. M. van der Heijde, *Arthritis Care Res.* **2010**, *62*, 624.
[40] J. Ptacek, R. E. Hawtin, D. Sun, B. Louie, E. Evensen, B. B. Mittleman, A. Cesano, G. Cavet, C. O. Bingham 3rd, S. S. Cofield, J. R. Curtis, M. I. Danila, C. Raman, R. A. Furie, M. C. Genovese, W. H. Robinson, M. C. Levesque, L. W. Moreland, P. A. Nigrovic, N. A. Shadick, J. R. O'Dell, G. M. Thiele, E. W. S. Clair, C. C. Striebich, M. B. Hale, H. Khalili, F. Batliwalla, C. Aranow, M. Mackay, B. Diamond, G. P. Nolan, P. K. Gregersen, S. L. Bridges Jr, *PLoS One* **2021**, *16*, e0244187.
[41] D. Sun, T. M. Nguyen, R. J. Allaway, J. Wang, V. Chung, T. V. Yu, M. Mason, I. Dimitrovsky, L. Ericson, H. Li, Y. Guan, A. Israel, A. Olar, B. A. Pataki, G. Stolovitzky, J. Guinney, P. S. Gulko, M. B. Frazier, J. C. Costello, J. Y. Chen, S. L. Bridges Jr, RA2 DREAM Challenge Community, *bioRxiv* **2021**, DOI 10.1101/2021.10.25.21265495.
[42] A. S. Chaudhari, C. M. Sandino, E. K. Cole, D. B. Larson, G. E. Gold, S. S. Vasanawala, M. P. Lungren, B. A. Hargreaves, C. P. Langlotz, *J. Magn. Reson. Imaging* **2021**, *54*, 357.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, **2009**.
[44] M. He, Z. Li, C. Liu, D. Shi, Z. Tan, *Asia Pac J Ophthalmol (Phila)* **2020**, *9*, 299.
[45] S. Budd, E. C. Robinson, B. Kainz, *Med. Image Anal.* **2021**, *71*, 102062.