Deep learning-based automatic-bone-destruction-evaluation system using contextual information from other joints

Kazuki Miyama (kazuki.miyama@human.ait.kyushu-u.ac.jp)
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Ryoma Bise
Department of Advanced Information Technology, Kyushu University
Satoshi Ikemura
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Kazuhiro Kai
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Masaya Kanahori
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Shinkichi Arisumi
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Taisuke Uchida
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Yasuharu Nakashima
Department of Orthopaedic Surgery, Graduate School of Medical Sciences, Kyushu University
Seiichi Uchida
Department of Advanced Information Technology, Kyushu University

Research Article

Keywords: Rheumatoid arthritis, Deep neural networks, modified Sharp/van der Heijde score, Automatic detection, Automatic classification

Posted Date: February 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1360101/v1

Version of Record: A version of this preprint was published at Arthritis Research & Therapy on October 3rd, 2022. See the published version at https://doi.org/10.1186/s13075-022-02914-7.

Abstract

Background: X-ray images are commonly used for assessing presence of bone destruction of rheumatoid arthritis. The purpose of this study is to propose an automatic-bone-destruction-evaluation system fully utilizing deep neural networks (DNN). This system detects all target joints of the modified Sharp/van der Heijde score (SHS) from a hand X-ray image. It then classifies every target joint as intact (SHS = 0) or non-intact (SHS ≥ 1).

Methods: We used 226 hand X-ray images of 40 rheumatoid arthritis patients. As for detection, we used a DNN model called DeepLabCut. As for classification, we built four classification models that classify the detected joint as intact or non-intact. First, each joint is independently classified; second, the same contralateral joint is compared; third, the same joint group (e.g., the proximal interphalangeal joints) of one hand are compared; fourth, the same joint group of both hands are compared. We evaluated DeepLabCut's detection performance and classification models' performances. The classification models' performances were compared to three orthopedic surgeons.

Results: Detection rates for all the target joints were 98.0% and 97.3% for erosion and joint space narrowing (JSN). Among the four classification models, the model that compares the same contralateral joint showed the best F-measure (0.70, 0.81) and area under the curve of the precision-recall curve (PR-AUC) (0.73, 0.85) regarding erosion and JSN. As for erosion, the F-measure and PR-AUC of this model were better than the best of the orthopedic surgeons.

Conclusions: The proposed system was useful. All the target joints were detected with high accuracy. The classification model that compared the same contralateral joint showed better performance than the orthopedic surgeons regarding erosion.

Introduction

Assessing presence of bone destruction is important for diagnosing rheumatoid arthritis (RA) [1,2]. In clinical settings, bone destruction is usually estimated as accurately as possible by observing X-ray images [3]. If the X-ray images reveal signs of bone destruction, the chance of an RA diagnosis is significantly increased, and early drug treatment is more likely to be suggested [4].

The modified Sharp/van der Heijde score (SHS) [5] is a commonly used metric for evaluating bone destruction by X-rays [6–8]. SHS has two assessment items: erosion and joint space narrowing (JSN). Erosion is assessed in 16 joints for each hand and wrist. The 16 assessed joints are the four proximal interphalangeal (PIP) joints, the interphalangeal joint of the thumb (IP), the five metacarpophalangeal (MCP) joints, the carpometacarpal (CMC) joint of the thumb, the multangular, the navicular, the lunate, the radius, and the ulna. JSN is assessed in 15 joints for each hand and wrist. The 15 assessed joints are the four PIP joints, the five MCP joints, the three CMC joints, the multangular-navicular joint, the capitate-navicular-lunate joint, and the radiocarpal joint. SHS for erosion has six grades from 0 to 5, and SHS for JSN has five grades from 0 to 4 [5].

Among the grades, the classification between 0 (intact) and the others (non-intact) is the most important task for early diagnosis of RA [6]. However, performing this binary classification by visual inspection is not easy, even for RA experts, for the three reasons described below.

First, binary classification of all joints by visual inspection is time-consuming [1,9–11]. Second, accurate and stable classification requires extensive practical experiences. Third, although each RA expert attempts the classification to the best of their ability, it is difficult to avoid intra- and inter-expert variability. These three problems make it difficult to perform binary classification in actual clinical practice.

The purpose of this study is to propose a system for automatically evaluating bone destruction (hereafter, "automatic-bone-destruction-evaluation system") by fully utilizing recent artificial-intelligence (AI) techniques, called deep neural networks (DNN). The system first detects all target joints from a single X-ray image of the hands. It then classifies every target joint as "intact" (SHS = 0) or "non-intact" (SHS ≥ 1).

The contributions of this study are twofold. First, the proposed system is the first that can detect all the target joints automatically. As for methods developed in previous studies [12–14], change in brightness is used for detecting the target joints. Due to the complex structure of some of the joints, the target joints were limited to the PIP, IP, and MCP joints [12–14]. To detect all the target joints, we introduced a DNN model—called "DeepLabCut"—that can detect various objects accurately by re-training it with different targets [15].

Second, the proposing system is the first that performs binary classification for each joint while utilizing "contextual" information from other joints. As for RA, bone destruction progresses bilaterally and symmetrically [1,3,16–18]. Therefore, comparing both hands is useful when reading X-ray images of RA patients [9,19–21]. Furthermore, the bone destruction of the same joint group, such as the PIP joints, tends to progress similarly [22,23]. These comparisons of the related joints are often used in diagnosing RA. We therefore propose a method that evaluates the target joints by using information concerning the relevant joints.

Materials And Methods

Dataset

This study was approved by the Institutional Review Board. We used 226 hand X-ray images from 40 patients who were diagnosed with RA and treated with medication from 2008 to 2019 in nine hospitals. Three orthopedic surgeons who treat RA gave the ground truth (GT) for each target joint as "intact" or "non-intact" by consensus. The number of intact and non-intact GTs for each joint are given as follows (intact, non-intact, total): in regard to erosion, the PIP-IP joints (1005, 125, 1130), the MCP joints (987, 143, 1130), the CMC joint of the thumb (172, 54, 226), the multangular (149, 77, 226), the navicular (96, 130, 226), the lunate (124, 102, 226), the radius (113, 113, 226), and the ulna (101, 125, 226); and in regard to JSN, the PIP joints (562, 342, 904), the MCP joints (876, 254, 1130), the CMC joints (328, 350, 678), the multangular-navicular joint (75, 151, 226), the capitate-navicular-lunate joint (84, 142, 226), and the

radiocarpal joint (58, 168, 226). Some joints are biased to be intact or non-intact, and this bias, called “class imbalance,” adversely affects prediction performance of the DNN [24]. To mitigate this class imbalance, we applied the data augmentation [25] as described below.

Method (overview)

Figure 1 overviews the proposed automatic-bone-destruction-evaluation system, which consists of three steps. First, a detection model detects the center point of the target joints (16 joints for erosion and 15 joints for JSN) from an inputted hand X-ray image (**Figure 1A**). Next, each joint image is cropped on the basis of the detected center point. The cropped image is then input into the classification model for binary classification (**Figure 1B**).

Training of detection model

As the detection model, DeepLabCut [15], which was proposed for detecting and tracking an animal’s joints in video images, was used. DeepLabCut estimates each key point (joint) position from an input image. This detection model has three advantages: first, it can be applied to various objects by re-training it with a different target; second, it can be trained with a few labeled training data; and third, it can detect joints accurately by learning the structure of the joints (e.g., mutual positional relationships). We thus used DeepLabCut for detecting the target joints.

In the training of DeepLabCut, 20 X-ray images were randomly selected from 226 X-rays and resized to 256 × 256 pixels. Then, for each of the 20 images, the center points of the target joints were annotated by an orthopedic surgeon. Finally, two DeepLabCut models were trained with those training images: one to detect the target joints for erosion (16 joints) and the other to detect the target joints for JSN (15 joints).

Models for classification

Several binary classification models were established, and their performances were compared via experiments. **Figure 2A** shows the baseline model, called the single-input single-output (SISO) model. The SISO model is the simplest model and classifies each detected joint independently. In other words, it does not utilize the information concerning other joints. As the backbone of the SISO model, a VGG16 [26], which has a typical convolutional neural network [27] structure and high object-recognition performance, was used. It is well-known that pre-training a neural network model with a large but non-target dataset boosts its performance. The VGG16 was therefore pre-trained by ImageNet [28] and then fine-tuned [29] by using (a limited number of) joint images.

In addition to the SISO model, three types of multiple-input multiple-output (MIMO) models were established. As discussed in the introduction, it is important to compare the same contralateral joint [9,19–21] or the same joint group [22]. From this viewpoint, the joint independence of the SISO model has room of improvement. The established MIMO models therefore utilize information about the same contralateral joint and the same joint group (**Figures 2B, 2C, and 2D**).

For designing the MIMO models, the same joint group in one hand are defined as follows (**Figure 2E**): for erosion, (1) the PIP and IP joints (PIP-IP joints), (2) the MCP joints, (3) the CMC joint of the thumb and multangular (CMC-M), (4) the wrist joints (the navicular, the lunate, the radius, and the ulna); and for JSN, (1) the PIP joints, (2) the MCP joints, (3) the CMC joints, (4) the wrist joints (the multangular-navicular joint, the capitate-navicular-lunate joint, and the radiocarpal joint). As for erosion, the navicular, lunate, radius, and ulna are grouped as wrist joints, and as for JSN, the multangular-navicular joint, capitate-navicular-lunate joint, and radiocarpal joint are grouped as wrist joints because the degree of bone destruction in these joints tends to be similar [22,23]. Given a set of joint images (the same joints of both hands or the same joint group) as inputs, the model simultaneously estimates the classes (“intact” or “non-intact”) of these multiple joints.

Figure 2B shows the *MIMO local* model, which can compare the same contralateral joint. This MIMO model receives inputs from the same joints of both hands and outputs “intact” or “non-intact” for each joint. Since the range of comparison is localized to the same joints of both hands, this model is referred to as the “MIMO local model.”

Figure 2C shows the *MIMO one-hand* model, which can compare multiple joints in the same joint group of one hand. This model simultaneously classifies each joint as “intact” or “non-intact” by using the information concerning the same joint group of one hand.

Figure 2D shows the *MIMO both-hands* model, which can compare the same joint group of both hands. This model may take advantages of both the MIMO local model and the MIMO one-hand model.

Training of classification models

The four classification models (SISO and three MIMOs) were trained under the following conditions. 40 patients were divided into eight subsets (five patients in each subset), eight-fold cross-validation was performed, where six subsets were used for training, one subset was used for validation, and one subset was used for testing. The classification models were trained by using the training images. The validation data was used to adjust hyperparameters and determine when to stop the training prematurely, namely, “early stopping,” which is used to improve the generalization of the test data [30]. The condition for early stopping validation loss does not decrease 10 times in a row. If this early stopping condition is not satisfied for 100 epochs, the training is terminated. We set the binary cross-entropy loss [31] for the SISO model and the sum of the binary cross-entropy losses over all outputs for the MIMO models.

Since image features among different joint groups differ significantly, different models were prepared for each group. As a result, for erosion, six SISO models and six MIMO local models were built (the PIP-IP joints, the MCP joints, the CMC-M, the proximal carpal bones [the navicular and the lunate], the radius, and the ulna), and for JSN, five SISO models and five MIMO local models were built (the PIP joints, the MCP joints, the CMC joints, the midcarpal joints [the multangular-navicular and the capitate-navicular-lunate joint], and the radiocarpal joint). Furthermore, for erosion, four MIMO one-hand models and four MIMO

both-hands models were built for each joint group (the PIP-IP joints, the MCP joints, the CMC-M, and the wrist joints), and for JSN, four MIMO one-hand models and four MIMO both-hands models were built for each joint group (the PIP joints, the MCP joints, the CMC joints, and the wrist joints).

To address the class imbalance, we applied data augmentation, which improves the performance of the DNN when there is a class imbalance or only a small amount of training data [25]. We applied data augmentation to the training and validation data (for each joint of each fold in the cross-validation). Specifically, we applied -3- to 3-degrees rotations, -5- to 5-pixels vertical and horizontal translations, and 0.97- to 1.03-times enlargement and/or reductions to each cropped joint image. The training data were augmented by these geometric perturbations until the total number of images was about 10,000 with no class imbalance. The validation data were also augmented until no class imbalance exists between intact and non-intact.

Procedure for evaluation of detection model

Detection performance of DeepLabCut was evaluated by using 206 test X-ray images. Since the X-ray images and hands have various scales and sizes, the X-ray images were first resized on the basis of the detected joints. More precisely, the X-ray images were resized so that the median lengths of the proximal phalanges in all images matched. Next, bounding boxes were formed around the detected center points. The box sizes are 250 × 250 pixels for the PIP, IP, MCP, and CMC-M joints, 500 × 300 pixels for the radius, and 300 × 300 pixels for the others. An orthopedic surgeon then checked whether the bounding box correctly contains the target joint. If not, the box is treated as an error and discarded from the later experiment. The performance of DeepLabCut was evaluated using the correct detection rate, that is, number of correct detections divided by the total number of joints.

Procedure for evaluation of classification models

The performance of the proposed four classification models was evaluated by using sensitivity, specificity, F-measure [32], and PR-AUC [33] with eight-fold cross-validation. F-measure and PR-AUC are important indicators of classification model performance when there is a class imbalance [34,35], as in this study. F-measure is the harmonic mean of sensitivity and precision, and PR-AUC is the curve of the area under the precision-recall curve, which is a plot of precision against sensitivity. Both F-measure and PR-AUC take values between 0 and 1 and become closer to 1 as performance improves.

Binary classification performance of the three orthopedic surgeons was also tested by using the same 226 X-rays. They evaluated each target joint for erosion and JSN as “intact” or “non-intact.” Compared to GT, their performances were measured using sensitivity, specificity, F-measure, and PR-AUC and compared with that of the classification models.

Results

Results (joint detection)

Table 1 shows the correct detection rates for each target joint. The detection rates for each joint are as follows (intact, non-intact, total): for erosion, the PIP-IP joints (99.5%, 90.8%, 98.5%), the MCP joints (99.6%, 83.1%, 97.5%), the CMC-M (99.3%, 93.4%, 97.6%), the wrist joints (100.0%, 96.3%, 98.1%), and all joints (99.6%, 92.9%, 98.0%); and for JSN, the PIP joints (99.0%, 91.6%, 96.2%), the MCP joints (98.2%, 86.8%, 95.6%), the CMC joints (99.7%, 98.7%, 99.2%), the wrist joints (100.0%, 99.8%, 99.8%), and all joints (98.9%, 95.2%, 97.3%). On the whole, all the target joints were detected with high accuracy. Intact joints (SHS = 0) were detected correctly in most cases. Detection performance was generally good in the case of non-intact joints (SHS ≥ 1), although detection rates for the PIP-IP and MCP joints tended to be a little low for both erosion and JSN.

Figure 3 shows examples of joint detection. In **Figure 3A**, joints with no or moderate bone destruction were successfully detected. In **Figure 3B,** in the case of severe bone destruction, false positives (yellow arrows) and false negatives (red arrows) for PIP and MCP joints were observed. On the contrary, the wrist joints could be detected accurately in terms of both erosion and JSN, even when severe bone destruction was present.

Results (classification)

Table 2 shows the binary classification performance of each classification model and each orthopedic surgeon in regard to three groups (the wrist joints, the others, and all joints [the wrist joints and the others]).For all joints, in the case of both erosion and JSN, the MIMO local model and the MIMO one-hand model outperformed the SISO model in terms of F-measure and PR-AUC. In addition, for all joints, the MIMO local model showed the best performance in terms of the following metrics: sensitivity of JSN (0.79), specificity of JSN (0.89), F-measure of erosion and JSN (0.70, 0.81), and PR-AUC of erosion and JSN (0.73, 0.85).

Furthermore, as for the F-measure and PR-AUC in the case of all joints, the MIMO local model showed better performance in regard to erosion than the best orthopedic surgeon and in regard to JSN than the average of the orthopedic surgeons. For all joints in regard to erosion and JSN, F-measure and PR-AUC were as follows (MIMO local model, average of the orthopedic surgeons, best of the orthopedic surgeons): F-measure (0.70, 0.58, 0.63) and PR-AUC (0.73, 0.67, 0.70) in regard to erosion; and F-measure (0.81, 0.80, 0.84) and PR-AUC (0.85, 0.84, 0.86) in regard to JSN.

Figure 4 shows examples of the visualization results of prediction by each classification model. In **Figure 4A**, the MIMO local model correctly classified the left hand’s navicular, lunate, and radius as non-intact, while the other models misrecognized them as intact. In regard to JSN, the MIMO local model correctly classified the right hand’s radiocarpal joint and the left hand’s CMC joints as non-intact, while the other models misrecognized them as intact (**Figure 4B**).

Discussion

The joint-detection performance of the proposed automatic-bone-destruction-evaluation system was very high for all the target joints. The proposed method for classifying a target joint as intact or non-intact shows high classification accuracy by comparing the same contralateral joint. In addition, the performance of the classification method in regard to erosion was higher than that of the orthopedic surgeons.

As for automatic detection, all the target joints could be detected with very high accuracy. Past reports [12,14] focused on changes in brightness values for automatic detection. These methods are effective when the joint space is well defined, such as in the PIP-IP and MCP joints, but it is ineffective in the case of anatomically complex structures such as the navicular and lunate [12,14]. DeepLabCut, which was used in this study, accurately detected all the target joints by learning the anatomical position and relationship between each target joint.

When ulnar drift occurred in the PIP and MCP joints (**Figure 3B**), detection by the proposed system tends to fail. This tendency can be explained by a change in the anatomical positional relationship between the proximal phalanges and the metacarpal bone. In contrast, the wrist joints could be appropriately detected, even though RA had progressed, because they had less anatomical deviation than the finger joints [36].

As for binary classification, the MIMO local model achieved the best performance among the four classification models. Comparing the same contralateral joint was more effective than comparing the same group joints. Previous studies [9,19–21] reported that comparing contralateral joints improves performance of reading X-ray images, and many rheumatologists have used this technique to diagnose bone destruction. Although the MIMO both-hands model also compares joints of both hands, its performance was not better than that of the MIMO local model and the MIMO one-hand model. We consider that increasing the number of input joints requires a combinatorial increase of training data, so it makes sufficient training difficult. It is thus important to select effective input images for a classification model.

In the non-intact cases, the MIMO local model and the MIMO one-hand model were effective in the wrist joints in the case of erosion and JSN. Bone destruction of the wrist joints progresses more symmetrically than the finger joints [16,37]; therefore, the MIMO local model was suitable for the wrist joints. The MIMO one-hand model was also effective in the wrist joints because the group of the wrist joints had a similar progression of bone destruction.

In the case of mild bone destruction in erosion, which is important for early diagnosis [38], the four classification models showed relatively good performance in regard to the wrist joints but not in regard to the PIP-IP, MCP, and CMC-M joints. It is even difficult for rheumatologists to accurately discriminate between intact bone and mild bone destruction [39]. The classification models may therefore have difficulty in learning features that discriminate between intact bone and mild bone destruction. A possible reason of the higher accuracy in regard to the wrist joints is that the wrist joints had less class imbalance. In regard to the PIP-IP, MCP, and CMC-M joints, prediction performance may be improved by increasing the number of non-intact cases.

The best-performing classification model was the MIMO local model; in particular, it shows better classification accuracy than the orthopedic surgeons for all joints in the case of erosion. As for JSN, general orthopedic surgeons are more familiar with evaluating it than erosion because they treat several diseases (such as osteoarthritis) for which evaluating JSN is important. Thus, their performance in regard to JSN was better than that in regard to erosion. Despite such a situation, the MIMO local model is slightly better than the average of the orthopedic surgeons in regard to all joints in the case of JSN. Erosion is more important than JSN in diagnosing RA [38]. Therefore, it is significant that the MIMO local model showed better performance than the orthopedic surgeons in the case of erosion. We thus conclude that the MIMO local model has a higher classification ability than the orthopedic surgeons in the case of erosion.

The limitations of this study are small number of data and lack of data about the feet. To address the limitation of the small number of data, we need to increase the number of data. More data may improve classification accuracy in the case of mild bone destruction in finger joints. A combination of X-ray evaluations of the hands and feet would be useful for early diagnosis of RA [40]. Accordingly, we will add foot data to enable evaluation of bone destruction in the feet by the proposed classification models.

In future work, we aim to improve classification performance by incorporating time-series information into the classification models, which is important for SHS scoring [9,41]. Using the same framework as the MIMO local model will make it is possible to incorporate time-series information.

Conclusion

In conclusion, the proposed automatic-bone-destruction-evaluation system was effective. As for automatic detection, all the target joints were detected with high accuracy. As for automatic binary classification, the proposed classification method, which could compare the same contralateral joint, showed generally good prediction performance for both erosion and JSN. In addition, the prediction performance of binary classification by the proposed method was better than that of the three orthopedic surgeons in regard to erosion.

Abbreviations

AI: artificial-intelligence; CMC: carpometacarpal; CMC-M; carpometacarpal joint of the thumb and multangular; DNN: deep neural networks; GT: ground truth; IP: interphalangeal; JSN: joint space narrowing; MCP: metacarpophalangeal; MIMO: multiple-input multiple-output; PIP: proximal interphalangeal; RA: rheumatoid arthritis; SHS; modified Sharp/van der Heijde score; SISO; single-input single-output

Declarations

Acknowledgment

The authors gratefully acknowledge Shota Harada, Kazuya Nishimura, and Kengo Araki for their great help in preparing the manuscript.

Authors' contributions

All authors contributed to the study conception and design.

KM, SI, KK, MK, SA, TU, and YN collected the data.

KM, RB, and SU wrote the manuscript, which was read and approved by all authors.

Funding

This study was supported by the Grants-in-Aid for Scientific Research of Japan Society for the Promotion of Science, Grant Number 19K09652.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author, KM, upon reasonable request.

Ethics approval and consent to participate

Ethical approvals for this study were obtained from the institutional review boards of all nine participating institutions.

Consent for publication

Not applicable.

Competing interests

The authors declare no conflicts of interest associated with this manuscript.

References

1. Salaffi F, Carotti M, Carlo M. Conventional radiography in rheumatoid arthritis: New scientific insights and practical application. Int J Clin Exp Med. 2016;9:17012–27.
2. Devauchelle Pensec V, Saraux A, Berthelot JM, Alapetite S, Chalès G, Le Henaff C, et al. Ability of hand radiographs to predict a further diagnosis of rheumatoid arthritis in patients with early arthritis. J Rheumatol. The Journal of Rheumatology; 2001;28:2603–7.
3. Drosos AA, Pelechas E, Voulgari PV. Conventional radiography of the hands and wrists in rheumatoid arthritis. What a rheumatologist should know and how to interpret the radiological findings. Rheumatol Int. Springer Science and Business Media LLC; 2019;39:1331–41.
4. McQueen FM. Imaging in early rheumatoid arthritis. Best Pract Res Clin Rheumatol. 2013;27:499–522.
5. van der Heijde DM, van Riel PL, Nuver-Zwart IH, Gribnau FW, vad de Putte LB. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet. 1989;1:1036–8.
6. Wen J, Liu J, Xin L, Wan L, Jiang H, Sun Y, et al. Effective factors on Sharp Score in patients with rheumatoid arthritis: a retrospective study. BMC Musculoskelet Disord. 2021;22:865.
7. Mochizuki T, Yano K, Ikari K, Hiroshima R, Sakuma Y, Momohara S. Correlation between hand bone mineral density and joint destruction in established rheumatoid arthritis. J Orthop. 2017;14:461–5.
8. Brown LE, Frits ML, Iannaccone CK, Weinblatt ME, Shadick NA, Liao KP. Clinical characteristics of RA patients with secondary SS and association with joint damage. Rheumatology. 2015;54:816–20.
9. van der Heijde DMFM. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillière's Clinical Rheumatology. 1996;10:435–53.
10. Brower AC. Use of the radiograph to measure the course of rheumatoid arthritis. Arthritis Rheum. Wiley; 1990;33:316–24.
11. Matsuno H, Yudoh K, Hanyu T, Kano S, Komatsubara Y, Matsubara T, et al. Quantitative assessment of hand radiographs of rheumatoid arthritis: interobserver variation in a multicenter radiographic study. J Orthop Sci. Elsevier BV; 2003;8:467–73.
12. Hirano T, Nishide M, Nonaka N, Seita J, Ebina K, Sakurada K, et al. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. Rheumatol Adv Pract. 2019;3:rkz047.
13. Nakatsu K, Morita K, Yagi N, Kobashi S. Finger Joint Detection Method in Hand X-ray Radiograph Images Using Statistical Shape Model and Support Vector Machine. 2020 International Symposium on Community-centric Systems (CcS). 2020. p. 1–5.
14. Morita K, Chan P, Nii M, Nakagawa N, Kobashi S. Finger Joint Detection Method for the Automatic Estimation of Rheumatoid Arthritis Progression Using Machine Learning. Available from: https://www.researchgate.net/publication/330478731
15. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21:1281–9.
16. Halla JT, Fallahi S, Hardin JG. Small joint involvement: a systematic roentgenographic study in rheumatoid arthritis. Ann Rheum Dis. 1986;45:327–30.
17. Zangger P, Keystone EC, Bogoch ER. Asymmetry of small joint involvement in rheumatoid arthritis: prevalence and tendency towards symmetry over time. Joint Bone Spine. 2005;72:241–7.
18. Sommer OJ, Kladosek A, Weiler V, Czembirek H, Boeck M, Stiskal M. Rheumatoid arthritis: a practical guide to state-of-the-art imaging, image interpretation, and clinical implications. Radiographics. 2005;25:381–98.
19. Ory PA. Interpreting radiographic data in rheumatoid arthritis. Ann Rheum Dis. 2003;62:597–604.
20. Fries JF, Bloch DA, Sharp JT, McShane DJ, Spitz P, Bluhm GB, et al. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum. 1986;29:1–9.

21. Ferrara R, Priolo F, Cammisa M, Bacarini L, Cerase A, Pasero G, et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the GRISAR study. Ann Rheum Dis. BMJ Publishing Group Ltd; 1997;56:608–12.
22. Hulsmans HM, Jacobs JW, van der Heijde DM, van Albada-Kuipers GA, Schenk Y, Bijlsma JW. The course of radiologic damage during the first six years of rheumatoid arthritis. Arthritis Rheum. 2000;43:1927–40.
23. Scott DL, Coulton BL, Popert AJ. Long term progression of joint damage in rheumatoid arthritis. Ann Rheum Dis. 1986;45:373–8.
24. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. Elsevier; 2018;106:249–59.
25. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. Springer; 2019;6:60.
26. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR [Internet]. 2015 [cited 2021 Oct 16]; Available from: https://www.semanticscholar.org/paper/eb42cf88027de515750f230b23b1a057dc782108
27. Li Z, Liu F, Yang W, Peng S, Zhou J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Trans Neural Netw Learn Syst [Internet]. ieeexplore.ieee.org; 2021; PP. Available from: http://dx.doi.org/10.1109/TNNLS.2021.3084827
28. DENG, J. ImageNet : A large-scale hierarchical image database. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit [Internet]. 2009 [cited 2021 Oct 16]; Available from: https://ci.nii.ac.jp/naid/10027363646
29. Guo Y, Shi H, Kumar A, Grauman K, Rosing T, Feris R. Spottune: transfer learning through adaptive fine-tuning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. openaccess.thecvf.com; 2019. p. 4805–14.
30. Ying X. An Overview of Overfitting and its Solutions. J Phys Conf Ser. IOP Publishing; 2019;1168:022022.
31. Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for image classification. Journal of Advanced Trends in Computer. researchgate.net; 2020; Available from: https://www.researchgate.net/profile/Vamsidhar-Yendapalli/publication/344854379_Binary_cross_entropy_with_deep_learning_technique_for_Image_classification/links/5f93eed692851c14bce1ac68/Bir cross-entropy-with-deep-learning-technique-for-Image-classification.pdf
32. Canbek G, Sagiroglu S, Temizel TT, Baykal N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. 2017 International Conference on Computer Science and Engineering (UBMK). ieeexplore.ieee.org; 2017. p. 821–6.
33. Keilwagen J, Grosse I, Grau J. Area under precision-recall curves for weighted and unweighted data. PLoS One. journals.plos.org; 2014;9:e92209.
34. Movahedi F, Padman R, Antaki J. Limitations of ROC on Imbalanced Data: Evaluation of LVAD Mortality Risk Scores. ArXiv [Internet]. 2020 [cited 2021 Sep 15]; Available from: https://www.semanticscholar.org/paper/3a83bb7335038801013f3805f09572c3f2f12280
35. Nan Y, Chai KM, Lee WS, Chieu HL. Optimizing F-measure: A Tale of Two Approaches [Internet]. arXiv [cs.LG]. 2012. Available from: http://arxiv.org/abs/1206.4625
36. Read GO, Solomon L, Biddulph S. Relationship between finger and wrist deformities in rheumatoid arthritis. Ann Rheum Dis. 1983;42:619–25.
37. Klarlund M, Ostergaard M, Jensen KE, Madsen JL, Skjødt H, Lorenzen I. Magnetic resonance imaging, radiography, and scintigraphy of the finger joints: one year follow up of patients with early arthritis. The TIRA Group. Ann Rheum Dis. 2000;59:521–8.
38. van der Heijde D. Erosions versus joint space narrowing in rheumatoid arthritis: what do we know? Ann Rheum Dis. 2011;70 Suppl 1:i116-8.
39. Guillemin F, Billot L, Boini S, Gerard N, Ødegaard S, Kvien TK. Reproducibility and sensitivity to change of 5 methods for scoring hand radiographic damage in patients with rheumatoid arthritis. J Rheumatol. 2005;32:778–86.
40. Visser H. Early diagnosis of rheumatoid arthritis. Best Pract Res Clin Rheumatol. Elsevier; 2005;19:55–72.
41. van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology. Oxford Academic; 1999;38:1213–20.

Tables

Table 1. Correct detection rates for test data.

Erosion	Intact	Non-intact	Total
PIP-IP	916/921 (99.5%)	99/109 (90.8%)	1015/1030 (98.5%)
MCP	896/900 (99.6%)	108/130 (83.1%)	1004/1030 (97.5%)
CMC-M	289/291 (99.3%)	113/121 (93.4%)	402/412 (97.6%)
Wrist	393/393 (100.0%)	415/431 (96.3%)	806/824 (98.1%)
All joints	2494/2505 (99.6%)	735/791 (92.9%)	3229/3296 (98.0%)
JSN			
PIP	508/513 (99.0%)	285/311 (91.6%)	793/824 (96.2%)
MCP	782/796 (98.2%)	203/234 (86.8%)	985/1030 (95.6%)
CMC	304/305 (99.7%)	309/313 (98.7%)	613/618 (99.2%)
Wrist	197/197 (100.0%)	420/421 (99.8%)	617/618 (99.8%)
All joints	1791/1811 (98.9%)	1217/1279 (95.2%)	3008/3090 (97.3%)

Correct detection rates for intact (SHS = 0) joints, non-intact (SHS ≥ 1) joints, and the total for each target joint are shown. The numbers represent the number of correct detection/total cases (correct detection rates %). In regard to erosion, “Wrist” represents the navicular, the lunate, the radius, and the ulna; in regard to JSN, “Wrist” represents the multangular-navicular, the capitate-navicular-lunate, and the radiocarpal joint.

Note: GT, ground truth; PIP, proximal interphalangeal; IP, interphalangeal; MCP, metacarpophalangeal; CMC-M, carpometacarpal joint of the thumb and multangular; JSN, joint space narrowing; CMC, carpometacarpal.

Table 2. Performance of binary classification.

PIP-IP+MCP+ CMC-M (erosion)	SISO	MIMO local	MIMO one-hand	MIMO both-hands	Orthopedic surgeon 3	Orthopedic surgeon 2	Orthopedic surgeon 3	Average of surgeons
Sensitivity	0.65	0.54	0.51	0.31	0.64	**0.94**	0.93	0.84
Specificity	0.88	0.93	0.93	**0.96**	0.92	0.57	0.78	0.76
F-measure	0.50	0.52	0.50	0.38	**0.57**	0.36	0.51	0.48
PR-AUC	0.55	0.54	0.53	0.44	0.59	0.58	**0.65**	0.61
Wrist (erosion)								
Sensitivity	0.78	0.85	0.84	0.77	0.46	**0.98**	0.88	0.77
Specificity	0.86	0.85	0.83	0.88	**0.95**	0.30	0.65	0.64
F-measure	0.81	**0.84**	0.83	0.81	0.61	0.70	0.77	0.70
PR-AUC	0.86	**0.88**	0.87	0.87	0.81	0.77	0.81	0.80
All joints (erosion)								
Sensitivity	0.73	0.72	0.70	0.58	0.57	**0.96**	0.87	0.80
Specificity	0.88	0.92	0.92	**0.95**	0.93	0.53	0.77	0.74
F-measure	0.66	**0.70**	0.69	0.65	0.61	0.51	0.63	0.58
PR-AUC	0.69	**0.73**	0.72	0.70	0.66	0.66	0.70	0.67
PIP+MCP+CMC (JSN)								
Sensitivity	0.75	0.74	0.73	0.67	0.55	**0.90**	0.84	0.76
Specificity	0.84	0.90	0.89	0.88	**0.93**	0.83	0.87	0.88
F-measure	0.72	0.76	0.74	0.70	0.65	**0.80**	0.79	0.74
PR-AUC	0.76	0.81	0.79	0.75	0.74	**0.82**	0.82	0.79
Wrist (JSN)								
Sensitivity	0.81	0.87	0.83	0.84	0.80	**0.98**	0.91	0.90
Specificity	0.79	0.78	0.80	0.81	**0.84**	0.67	**0.84**	0.78
F-measure	0.85	0.88	0.86	0.87	0.85	0.91	**0.91**	0.89
PR-AUC	0.91	0.93	0.92	0.93	0.91	0.92	**0.94**	0.92
All joints (JSN)								
Sensitivity	0.77	0.79	0.77	0.74	0.62	**0.93**	0.86	0.80
Specificity	0.83	0.89	0.88	0.87	**0.92**	0.81	0.86	0.87
F-measure	0.76	0.81	0.79	0.76	0.71	**0.84**	0.83	0.80
PR-AUC	0.81	0.85	0.83	0.82	0.81	**0.86**	0.86	0.84

Performance of each classification model and orthopedic surgeons in regard to erosion and JSN are shown. In regard to erosion, “Wrist” represents the navicular, the lunate, the radius, and the ulna; in regard to JSN, “Wrist” represents the multangular-navicular, the capitate-navicular-lunate, and the radiocarpal joint.

Note: SISO, single-input single-output model; MIMO, multiple-input multiple-output; JSN, joint space narrowing; PR-AUC, precision-recall area under the curve. PIP, proximal interphalangeal; IP, interphalangeal; MCP, metacarpophalangeal; CMC-M, carpometacarpal joint of thumb and multangular; CMC, carpometacarpal.

Figures

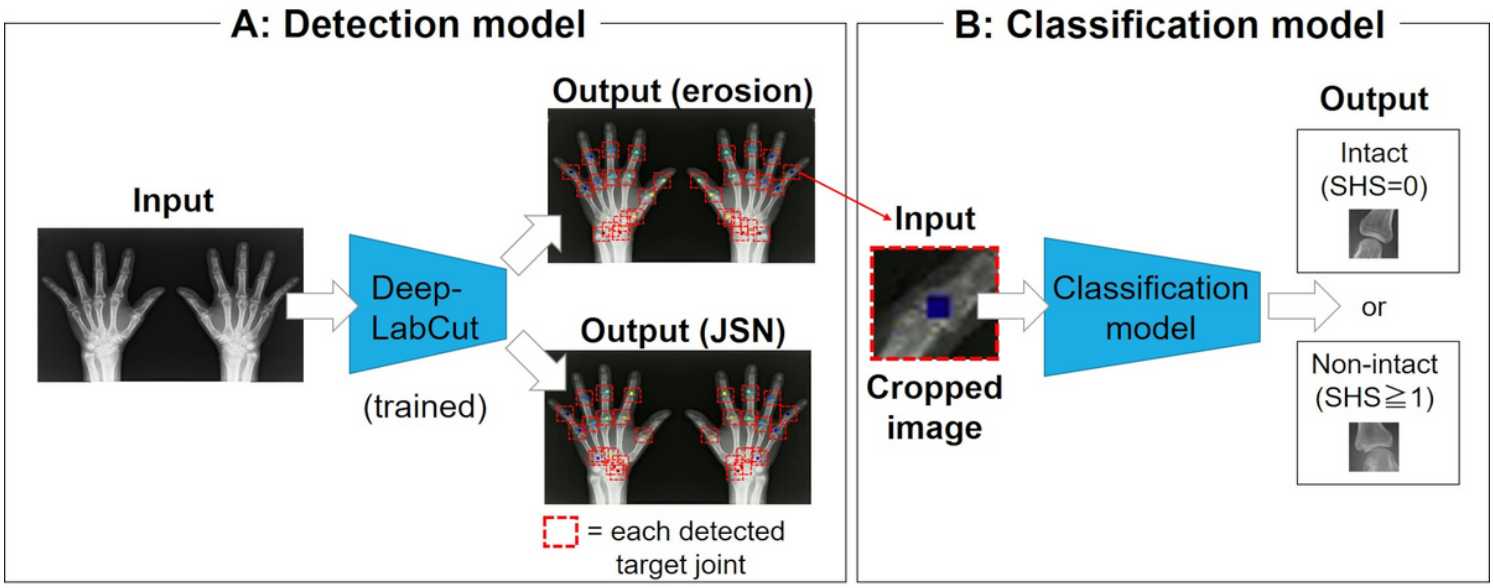

Figure 1

Overview of the proposed automatic-bone-destruction-evaluation system. **(A)** Input an X-ray image of hands into the detection model (DeepLabCut). DeepLabCut detects the center point of the evaluation joints of the SHS (target joints): 16 joints for erosion (the 4 PIP joints, the IP joint of the thumb, the 5 MCP joints, the CMC joint of the thumb, the multangular, the navicular, the lunate, the radius, and the ulna); 15 joints for JSN (the 4 PIP joints, the 5 MCP joints, the 3 CMC joints, the multangular-navicular joint, the capitate-navicular-lunate joint, and the radiocarpal joint). Each point indicates the detected center of the target joints. Each joint image was cropped (**Red bounding box**) according to the detected center point. **(B)** Each cropped image was input into the classification model, which outputs whether the input image is intact (SHS = 0) or non-intact (SHS ≥ 1).

Note: SHS, Sharp/van der Heijde score; PIP, proximal interphalangeal; IP, interphalangeal; MCP, metacarpophalangeal; CMC, carpometacarpal; JSN, joint space narrowing.

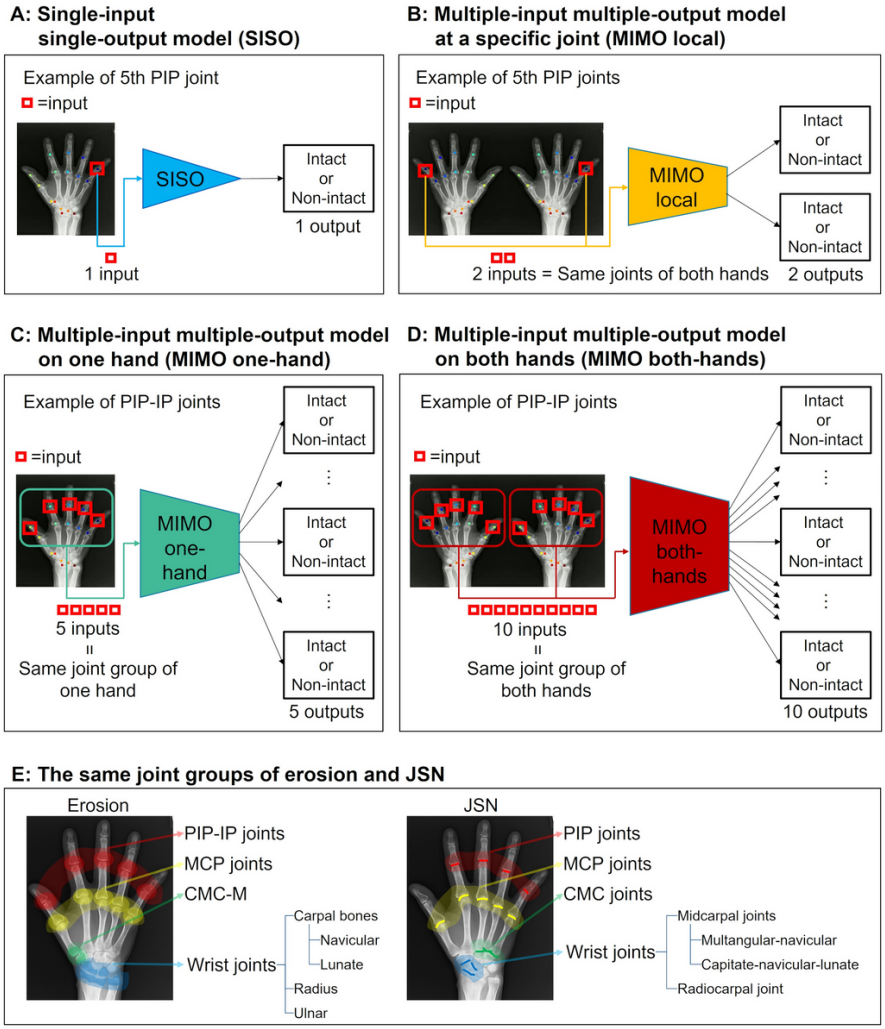

Figure 2

Overview of the binary classification model and the same joint groups. We developed four classification models: **(A)** a most-basic classification model that independently classifies each joint as intact (SHS = 0) or non-intact (SHS ≥ 1) (single-input single-output model). **(B)** a classification model that inputs the same joints of both hands and outputs intact or non-intact, respectively (multiple-input multiple-output model at a specific joint). **(C)** a classification model that receives inputs of the same joint group of one hand and outputs whether they are intact or non-intact, respectively (multiple-input multiple-output model on the one hand). **(D)** a classification model that receives inputs of the same joint group of both hands and outputs whether they are intact or non-intact, respectively (multiple-input multiple-output model on both hands). (**E**) shows the same joint groups of both hands in regard to erosion and JSN.

Note PIP, proximal interphalangeal; IP, interphalangeal; MCP, metacarpophalangeal; CMC-M, carpometacarpal joint of the thumb and multangular; CMC, carpometacarpal.

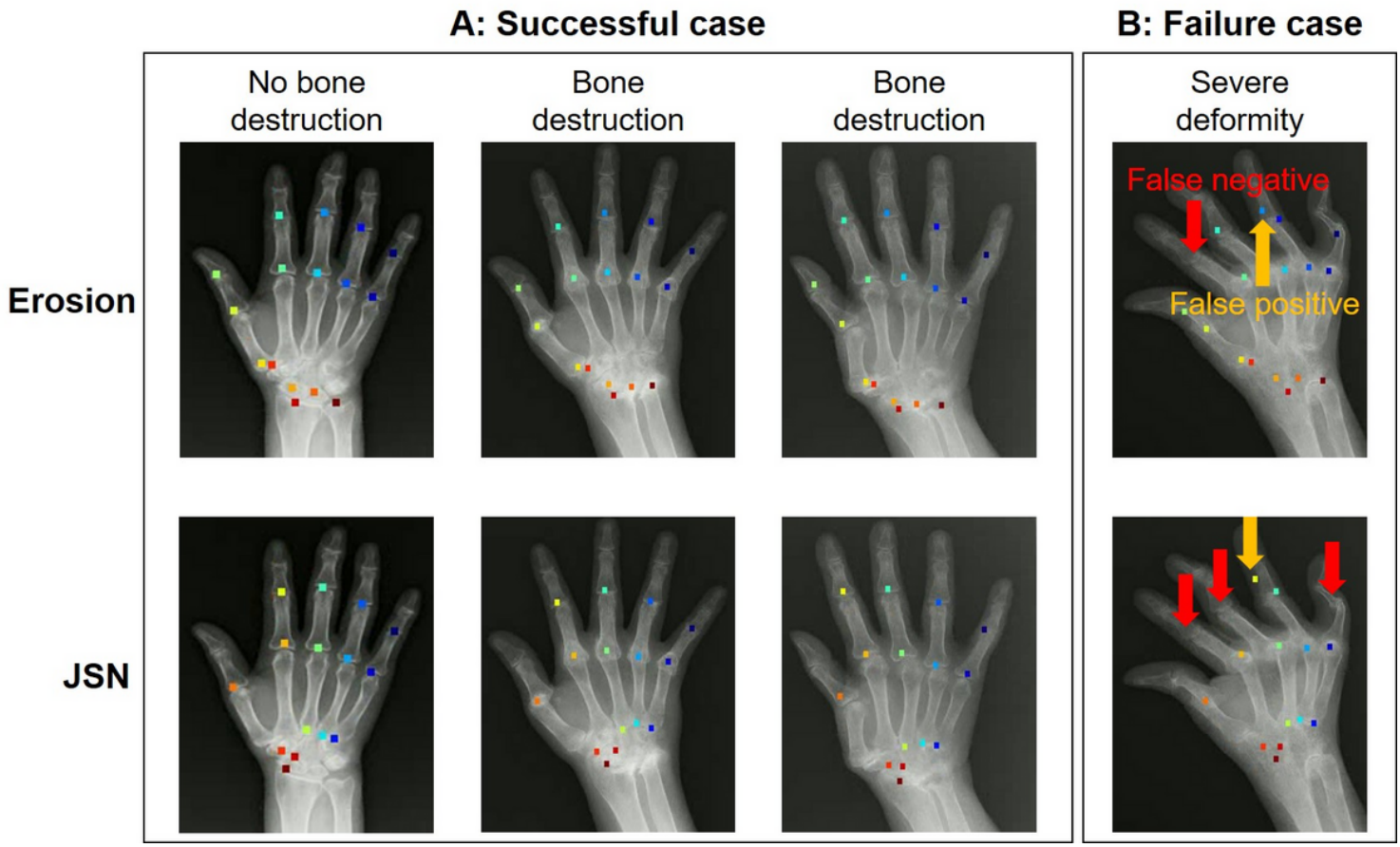

Figure 3

Detection results by DeepLabCut. Each point means the detected center point of the target joints, and each color corresponds to each joint. (**A**) DeepLabCut could detect the joints with high accuracy in the case with or without moderate bone destruction. **(B)** Detection failed in cases of severe bone destruction. The yellow arrow shows DeepLabCut detected the wrong area (**False positive**). The red arrow shows DeepLabCut could not detect the target joints (**False negative**).

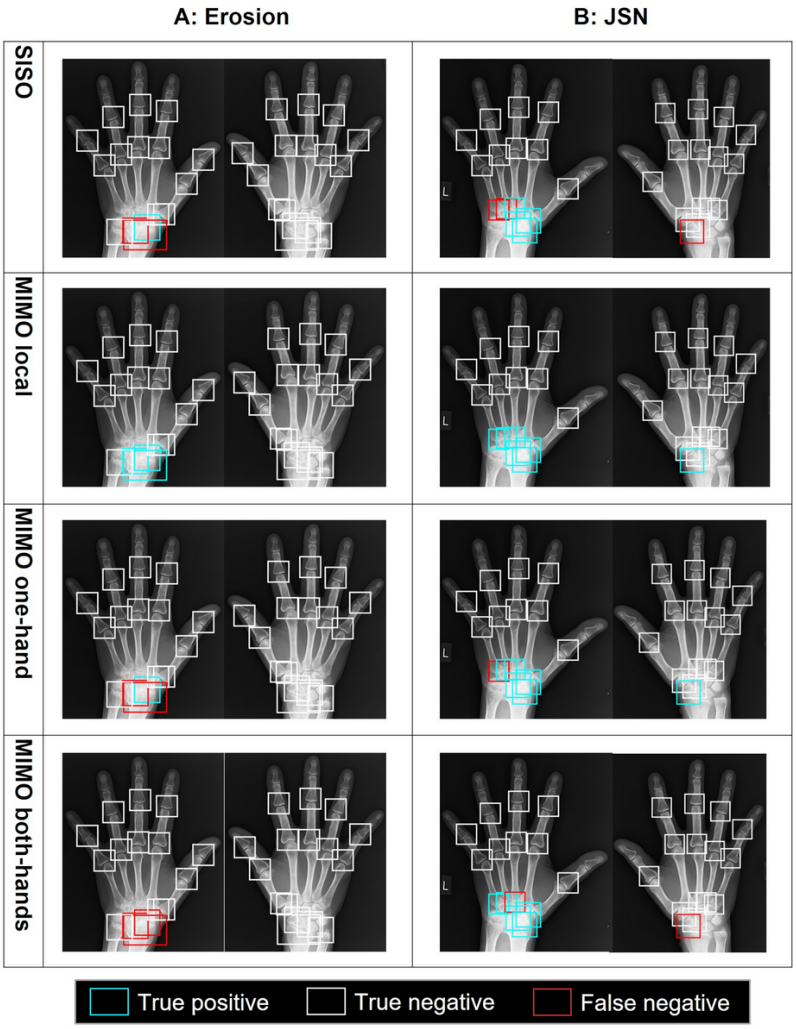

Figure 4

Visualization results of prediction by each classification model of target joints in X-ray images. Each bounding box represents the detected target joints. The color of the bounding box indicates how the classification results compare to ground truth (GT). Blue indicates GT is non-intact, and prediction is non-intact (**True positive**). White indicates GT is intact, and prediction is intact (**True negative**). Red indicates GT is non-intact, and prediction is intact (**False negative**). **(A)** shows the results for erosion and **(B)** shows the results for JSN.

Note: GT, ground truth; JSN, joint space narrowing.