THERAPEUTIC ADVANCES in
Musculoskeletal Disease

Original Research

An optimal deep learning model for the scoring of radiographic damage in patients with ankylosing spondylitis

Ther Adv Musculoskelet Dis

2024, Vol. 16: 1–11

DOI: 10.1177/1759720X241285973

Yen-Ju Chen, Der-Yuan Chen, Haw-Chang Lan, An-Chih Huang, Yi-Hsing Chen, Wen-Nan Huang and Hsin-Hua Chen

Abstract

Background: Detecting vertebral structural damage in patients with ankylosing spondylitis (AS) is crucial for understanding disease progression and in research settings.

Objectives: This study aimed to use deep learning to score the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS) using lateral X-ray images of the cervical and lumbar spine in patients with AS in Asian populations.

Design: A deep learning model was developed to automate the scoring of mSASSS based on X-ray images.

Methods: We enrolled patients with AS at a tertiary medical center in Taiwan from August 1, 2001 to December 30, 2020. A localization module was used to locate the vertebral bodies in the images of the cervical and lumbar spine. Images were then extracted from these localized points and fed into a classification module to determine whether common lesions of AS were present. The scores of each localized point were calculated based on the presence of these lesions and summed to obtain the total mSASSS score. The performance of the model was evaluated on both validation set and testing set.

Results: This study reviewed X-ray image data from 554 patients diagnosed with AS, which were then annotated by 3 medical experts for structural changes. The accuracy for judging various structural changes in the validation set ranged from 0.886 to 0.985, whereas the accuracy for scoring the single vertebral corner in the test set was 0.865.

Conclusion: This study demonstrated a well-trained deep learning model of mSASSS scoring for detecting the vertebral structural damage in patients with AS at an accuracy rate of 86.5%. This artificial intelligence model would provide real-time mSASSS assessment for physicians to help better assist in radiographic status evaluation with minimal human errors. Furthermore, it can assist in a research setting by offering a consistent and objective method of scoring, which could enhance the reproducibility and reliability of clinical studies.

Keywords: ankylosing spondylitis, artificial intelligence, deep learning, mSASSS, radiographic damage

Received: 31 October 2023; revised manuscript accepted: 5 September 2024.

Correspondence to:
Hsin-Hua Chen
Division of Allergy, Immunology and Rheumatology, Department of Internal Medicine, Taichung Veterans General Hospital, 1650 Taiwan Boulevard Sect. 4, Taichung 40705, Taiwan

Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung, Taiwan

School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

Department of Business Administration, Ling Tung University, Taichung, Taiwan

Department of Industrial Engineering and Enterprise Information, Tunghai University, Taichung, Taiwan

Institute of Biomedical Science and Rong Hsing Research Center for Translational Medicine, Big Data Center, National Chung Hsing University, Taichung, Taiwan

Institute of Public Health and Community Medicine Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan
shc5555@hotmail.com

Yen-Ju Chen
Division of Allergy, Immunology and Rheumatology, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan

Introduction

Ankylosing spondylitis (AS) is a chronic inflammatory rheumatic disease and a subset of axial spondyloarthritis (axSpA).[1] AS predominantly involves inflammation of the axial skeleton, including the spine and sacroiliac joints, as well as sometimes the peripheral joints.[2] The axial structural change seen in patients with AS may begin at a young age, and progress from damage in sacroiliac joints to the whole spine, leading to the impairment of spine motility.[3] There are many instruments available for evaluating the structural

Institute of Clinical Medicine, National Yang Ming University, Taipei, Taiwan

Department of Medical Research, Taichung Veterans General Hospital, Taichung, Taiwan

Der-Yuan Chen
Institute of Medicine, Chung Shan Medical University Hospital, Taichung, Taiwan

Rheumatology and Immunology Center, China Medical University Hospital, Taichung, Taiwan

College of Medicine, China Medical University, Taichung, Taiwan

Haw-Chang Lan
Department of Diagnostic Radiology, Tungs' Taichung MetroHarbor Hospital, Taichung, Taiwan

An-Chih Huang
Advanced Tech BU, Acer Incorporated, New Taipei City, Taiwan

Yi-Hsing Chen
Division of Allergy, Immunology and Rheumatology, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan

Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung, Taiwan

School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

Wen-Nan Huang
Division of Allergy, Immunology and Rheumatology, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan

Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung, Taiwan

School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

Department of Business Administration, Ling Tung University, Taichung, Taiwan

change of AS, including conventional radiographs, computed tomography, and magnetic resonance imaging. Although conventional radiographs are less sensitive, they remain an efficient, less time-consuming and the most widely used method for the detection of structural change in AS.[4] Several composite methods for assessing radiographic damage in AS using conventional radiographs have been proposed, including the Stoke Ankylosing Spondylitis Spinal Score (SASSS), the Bath Ankylosing Spondylitis Radiology Index, the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS), and the Radiographic Ankylosing Spondylitis Spinal Score (RASSS).[5–7] Establishing a method for assessing radiographic destruction is crucial in the evaluation of drug response for the purpose of preventing the progression of structural damage in patients with AS. Previous studies which compared multiple composite methods revealed that the mSASSS is the most valid and feasible tool for scoring radiographic damage in AS, including early axSpA.[8,9] One review article also indicates that the mSASSS is a reliable tool for determining the progression of radiographic damage in AS, which is associated with the worsening of clinical manifestations and physical function in patients with AS.[10] However, it is essential to acknowledge the limitations of mSASSS, including interobserver and intraobserver variability and reliability.[11]

mSASSS is a modification of SASSS, created by adding an additional scoring category for the cervical spine, thus making a more clear definition of squaring.[5] The totals taken from mSASSSs, which count the sum of the lumbar and cervical spine lesion scores, range from 0 to 72, with a score of 0 indicating no radiographic change, a score of 1 for erosion, squaring, or sclerosis, a score of 2 for syndesmophytes, and a score of 3 for bridging syndesmophytes or ankylosis. Since there are 5 damaged parts that need to be identified in 12 spine columns, scoring with mSASSS is time-consuming and adds to the workload of radiologists and rheumatologists.[12,13] This is not only labor-intensive but also subject to operative error. Consequently, obtaining help from artificial intelligence (AI) is urgent, not only for clinicians but also for clinical researchers, to better assist in the interpretation of radiographs and automatically give rise to mSASSSs in patients with AS. Currently, there are several clinical trials that support the efficacy of certain biologic agents, such as certolizumab, secukinumab, and ixekizumab, in delaying or even improving radiographic progression in patients with AS.[14–16] Slowing radiographic progression will lead to reduced disability, improved quality of life, and delayed disease complications in patients with AS. Rapid mSASSS calculation is therefore vital for clinicians to monitor disease progression and prompt treatments of adequate biologic agents to slow radiographic progression in patients with AS. Moreover, mSASSS is also utilized across numerous clinical trials to standardize radiographic changes in patients with AS, and AI-based mSASSS calculation can mitigate individual discrepancies in scoring, helping researchers accurately assess disease progression and treatment efficacy in clinical trials.

Recently, one study in Korea demonstrated a deep learning model for scoring the corner of the vertebrae in patients with AS using mSASSSs, with the accuracy, sensitivity, and specificity of one corner being 0.91604, 0.80288, and 0.94244, respectively.[17] However, the characteristics of different spinal structural lesions counted by one score are not differentiated, and subjects having spinal deformities or foreign body implantation were not included in the Korea cohort, thus making clinical application difficult. There is currently no appropriate real-world model for scoring structural damage in mSASSSs using machine deep learning and AI for patients with AS. Thus, our study aimed to develop an optimal model for the auto-calculation of mSASSSs using deep learning in patients with AS in Asian populations for assistance in clinical decision-making.

Methods

Enrollment of participants and data collection

We identified patients with AS from a tertiary medical center in Taiwan from August 1, 2001 to December 30, 2020. AS was diagnosed clinically and excluded if not fulfill the modified New York criteria or ASAS (Assessment of SpondyloArthritis International Society) classification criteria.[18–20] Radiographs were reviewed and collected in DICOM format. If X-ray images of the lateral view of cervical spines and lumbar spines were seen within 3 months, they would be defined as one set of images. In total, 1056 sets

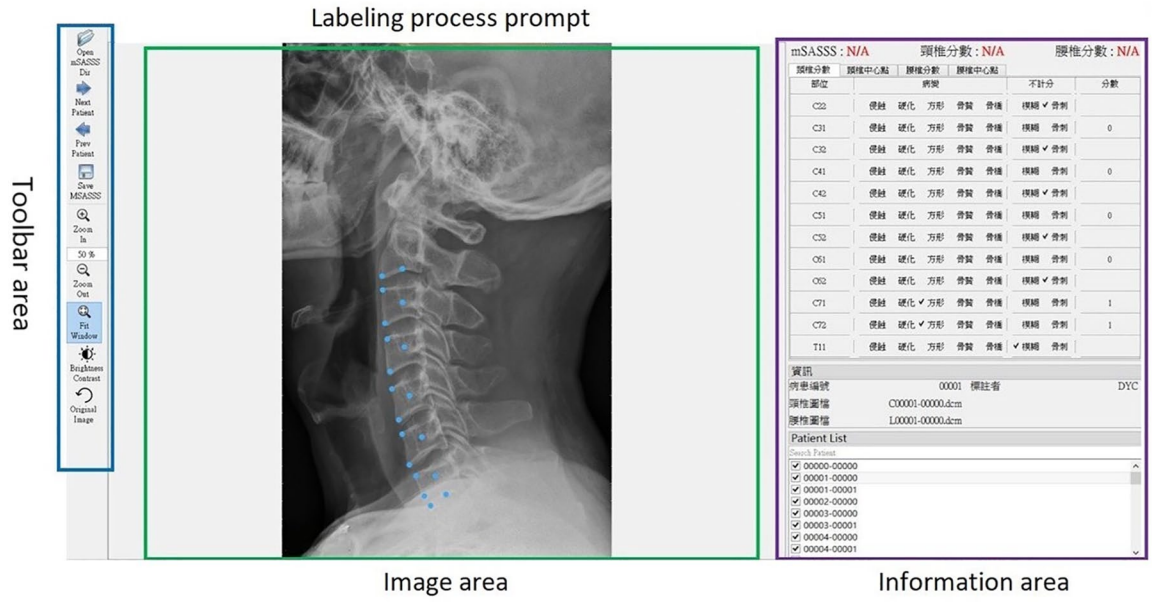

Figure 1. mSASSS labeling tool in patients with AS.
AS, ankylosing spondylitis; mSASSS, the modified Stoke Ankylosing Spondylitis Spinal Score.

of X-ray images from 554 patients with AS were collected.

The datasets of X-ray images were divided into three categories: the training set, the validation set, and the testing set. We randomly selected and categorized 20% of the datasets, resulting in 211 images being used as the testing set. When one set of a patient was selected into the testing set, other data belonging to this patient would also be selected into the testing set to ensure that the patients placed into the testing set were not grouped into the training and validation sets. The remaining 845 sets of datasets were used for the training and validation sets for the purpose of AI training.

Data labeling process

The lesions noted on the radiographs for cervical and lumbar spines were labeled by two rheumatologists and one radiologist. The labeling tool is shown in Figure 1. In mSASSS, there were 24 anterior corners of the vertebra to be labeled, including 12 anterior corners for cervical spines and 12 anterior corners for lumber spines. For each anterior corner, an annotator needed to be present to label the lesions. The label items for each anterior corner were: no abnormality (score 0), erosion (score 1), sclerosis (score 1), squaring (score 1), syndesmophyte (score 2), total bony bridging at each site (score 3), osteophytes (corner not considered for scoring), others such as non-marginal syndesmophytes, intra-discal ankylosis, intervertebral ossification, and paravertebral ossification (corner not considered for scoring), not clearly visible or blurry (corner not considered for scoring).[5,21] Notably, the third cervical vertebra is not scored for squaring in mSASSS calculation due to its naturally straight shape on the lateral surface.[5] When discrepancies arose among physicians, consensus meetings were held to reach a decision based on the majority of opinion.

AI models construction and imbalanced data correction

Two types of AI models, localization model and classification model, were used in algorithm construction (Figure 2). The localization model was executed in order to find the position of each vertebral body, including the upper anterior corner, lower anterior corner, and center point of each vertebra. The AI architectures including U-Net-based segmentation, YOLO, and PoseNet were considered for localization initially. However, it was necessary to delineate the boundaries of all vertebrae in U-Net-based segmentation, which increased labeling time and labeling errors for physicians. For YOLO architecture, it could only frame a square area and once the vertebrae were

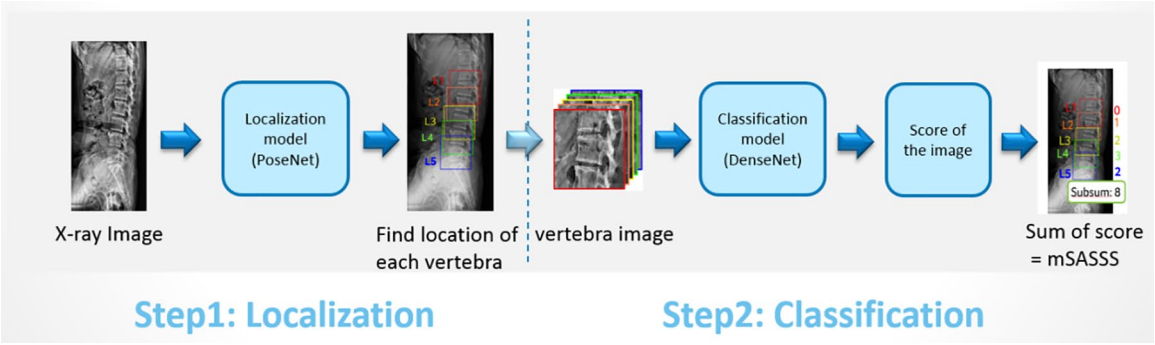

Figure 2. The AI model construction for mSASSS scoring in patients with AS.
AI, artificial intelligence; AS, ankylosing spondylitis; mSASSS, the modified Stoke Ankylosing Spondylitis Spinal Score.

not arranged in a straight and orderly manner, the squared frame might include nearby vertebrae. We finally selected the PoseNet-based model as the localization model, which was good for pose estimation in previous studies.[22,23] We further cut out small images from the cervical and lumbar X-ray images based on the above point information. These small images were sent to a classification model to check the corners for any lesions. Next, we calculated the score for each corner based on the lesion information, and then added up all the scores to develop mSASSS scores. The DenseNet-121-based model was used as the classification model.[24,25]

To enhance the efficacy of AI in identifying lesions, considering that the number of positive samples varies across different lesions, training all lesions with the same training and validation set does not yield optimal training outcomes for individual lesions. Therefore, each lesion identification AI model is equipped with its dedicated training and validation datasets, totaling eight sets. This approach aims for precise AI training, enabling effective identification of various lesions. Specifically, positive samples from each lesion were randomly selected and divided into the training set and the validation set at the ratio of 8:2. Moreover, due to the severe imbalance between positive and negative samples, if unaddressed, AI may tend to predict negative outcome after learning from an excess of negative samples, thus reducing prediction accuracy. To tackle this issue, the negative samples were randomly deleted, with the ratio of positive to negative samples ultimately being 1:3. This aims to achieve more balanced training conditions, thereby enhancing the AI's predictive accuracy.

Additionally, we took certain measures to increase the number of positive samples. Since the number of positive samples of an image was not equal to the number of lesion sites, for example, a bony bridge included the upper and lower corners of the adjacent vertebrae, which implied that the number of positive samples of bridges was only at half the number of lesion sites. To increase the number of positive samples, we vertically flipped the positive images and added them to the dataset; after that, we performed oversampling for some lesions to increase the chances of the AI model being trained on only positive samples. The DenseNet-121 model, the densely connected convolutional networks, was implemented due to the relatively small number of positive samples for most lesions.[24] Data augmentation techniques, such as rotation, width shift, height shift, and zoom, were also used to train the model. Furthermore, due to the low number of positive samples for sclerosis and squaring, sclerosis lesions seen in the cervical and lumbar images were merged together, as were the squaring lesions in the cervical and lumbar images. For erosion, there were too few positive samples, and combining cervical and lumbar spine images was not enough to train a model that yielded results.

Ethics statement

This study was approved by the Institutional Review Board of Taichung Veterans General Hospital, Taiwan (IRB No. CE20295A) and waived requirement for informed consent because patient data were anonymized before analysis.

The reporting of this study conforms to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement.[26]

Results

Demographic characteristics and specific lesions on radiographs in patients with AS

Table 1 shows the characteristics of patients with AS. The mean age at AS diagnosis was 26.5 ± 10.7 years, and most of the participants were men (76.4%). The numbers for the labeling of each lesion after the majority decision are shown in Table 2. Bony bridges (13.58%) were the most observed lesions seen in cervical spines, followed by osteophytes (6.38%), squaring (5.32%), and syndesmophytes (4.48%). As for lumbar spines, bony bridges (16.63%) were also the most seen, followed by osteophytes (6.83%), sclerosis (4.86%), and squaring (4.61%). The interobserver agreement was appropriate, ranging from 84.0% of squaring lesion to 99.2% of erosion lesion, as shown in Table 3. The interobserver agreement was defined as the number of agreements between the independent observers and divided by the total number of agreements plus disagreements.

AI model in the validation set

The training results from the validation set are summarized in Table 4, and the accuracy for judging various structural lesions in the validation set ranged from 0.886 to 0.985. The accuracy of bony bridge for cervical spine and lumbar spine was 0.985 and 0.974, respectively. The accuracy of syndesmophyte for cervical spine and lumbar spine was 0.9167 and 0.8936, respectively, whereas the accuracy of sclerosis and squaring for both cervical and lumbar spine was 0.916 and 0.886, respectively. Owing to a shortage of positive samples for erosions, it was difficult to obtain effective results using the deep learning model.

AI model in the testing set

Table 5 shows the confusion matrix on the testing set. In this section, the results are compared between the scores predicted by the AI model and those labeled by the rheumatologist, with a score range of 0–3 for each individual joint assessment. The overall accuracy of mSASSS site scores is 86.5%, where the accuracy of a score of 0 is 88.6% 1 is 67.7%, 2 is 73.8%, and 3 is 89.9%. It should be noted that the accuracy of a score of 2 is slightly lower than the results separately calculated for that of score 3. This is because a very

Table 1. Baseline characteristics of patients with AS.

Demography	*N*	%
Age at AS diagnosis (years), mean ± SD	26.5	10.7
Age of mSASSS score (years), mean ± SD	41.1	12.7
Gender, male	423	76.4
Clinical manifestations		
Uveitis	134	24.2
Psoriasis	41	7.4
Inflammatory bowel disease	3	0.5
Peripheral arthritis	102	18.4
Enthesitis	92	16.6
Dactylitis	8	1.4
Disease activities		
ESR (mm/h), mean ± SD	12.4	14
hsCRP (mg/dL), mean ± SD	0.5	0.9
ASDAS-ESR, mean ± SD	1.7	0.9
0 ⩽ ASDAS-ESR < 1.3	184	33.2
1.3 ⩽ ASDAS-ESR < 2.1	202	36.5
2.1 ⩽ ASDAS-ESR ⩽ 3.5	147	26.5
3.5 < ASDAS-ESR	21	3.8
ASDAS-CRP, mean ± SD	1.7	0.9
0 ⩽ ASDAS-CRP < 1.3	198	35.7
1.3 ⩽ ASDAS-CRP < 2.1	188	33.9
2.1 ⩽ ASDAS-CRP ⩽ 3.5	144	26
3.5 < ASDAS-CRP	24	4.3
BASDAI, mean ± SD	2.5	1.8
0 ⩽ BASDAI < 3	348	62.8
3 ⩽ BASDAI < 4	88	15.9
4 ⩽ BASDAI < 6	91	16.4
6 ⩽ BASDAI	27	4.9

(Continued)

Table 1. (Continued)

Demography	*N*	%
Comorbidities		
Hypertension	111	20
Diabetes mellitus	46	8.3
Hyperlipidemia	71	12.8
Chronic kidney disease	9	1.6
Coronary artery disease	17	3.1
Fracture, any sites	54	9.7
Post total hip replacement	18	3.2
Post total knee replacement	3	0.5
Family history		
AS, the first generation	102	18.4
AS, the second generation	152	27.4
Psoriasis	21	3.8
Psoriatic arthritis	4	0.7
Uveitis	27	4.9
Inflammatory bowel disease	2	0.2
Rheumatoid arthritis	34	6.1
Systemic lupus erythematosus	17	3.1
Sjogren's syndrome	10	1.8

AS, ankylosing spondylitis; ASDAS-CRP, The Ankylosing Spondylitis Disease Activity Score with C-reactive protein; ASDAS-ESR, The Ankylosing Spondylitis Disease Activity Score with ESR; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; ESR, erythrocyte sedimentation rate; mSASSS, modified Stoke Ankylosing Spondylitis Spinal Score; SD, standard deviation.

small number of sites were judged by the AI model to have structural changes of both bridge and syndesmophyte simultaneously. In such cases, only a score of 3 is shown, resulting in slightly fewer sites being counted as having a score of 2. The overall accuracy rate of score judgments in this test set is 86.5%.

Discussion

This study demonstrated the use of a deep learning model for scoring mSASSS to detect vertebral structural damage in patients with AS. The accuracy for judging various structural lesions in the validation set ranged from 0.886 to 0.985, with the accuracy of all vertebral corners scoring from the perspective of the end user being 86.5%, a result expected to have solid clinical practicability.

The AI model accuracy of lesions in the validation set for all types of lesions was slightly better for the cervical spine in comparison to that of the lumbar spine, which could be attributed to the anatomical differences and the clarity of lesion manifestation in cervical X-rays. The cervical spine's relatively straightforward structure may allow for more precise lesion identification. Additionally, the lumbar spine often has more osteophytes and may have undergone surgeries, which can complicate lesion scoring as these factors are not counted.

Owing to no extensively large numbers of lesions being seen in our mSASSS prediction model, DenseNet, a set of small-to-medium-sized networks, was selected as the basic AI model for feasibility tests so as to avoid overfitting. Furthermore, because there were upper and lower corners in each patch image, and each corner may possess multiple lesions, we used the AI model of multi-task multi-label architecture to initially evaluate the structural damage; however, the resulting accuracy was found to be unsatisfactory. The reason for this could be that there was a shortage of certain structural lesion data, such as that regarding erosions, as well as that there being data imbalance between each lesion. As a result, we went ahead and further used one label architecture to assess one single lesion at a time, which contained a basic Convolutional Neural Network (CNN) Model, Global Average Pooling layer (GAP), and had the task to the upper and lower corners. CNN is a deep learning model for image processing, using convolutional layers to extract features and fully connected layers for tasks like classification. Although all structurally damaged lesions could not be evaluated at once, this would up markedly increasing the recognition accuracy by avoiding any data imbalance between each lesion in multi-task multi-label architecture. Of note, the number of lesion images was critical toward the achievement of AI model training. Due to an increasing number of lesion labels and the interactive training that occurred via a combination of cervical and lumbar spine patch images through one label architecture, specific lesions such as sclerosis and squaring were successfully

Table 2. Labeling for spinal lesions in cervical and lumbar spines in patients with AS after majority decision.

Lesions	Cervical spines		Lumbar spines	
	N	%	*N*	%
Erosion	5	0.04	23	0.18
Sclerosis	96	0.76	616	4.86
Squaring	674	5.32	584	4.61
Syndesmophyte	568	4.48	520	4.10
Bony bridge	1721	13.58	2107	16.63
Osteophyte	808	6.38	797	6.83

AS, ankylosing spondylitis.

trained. Nonetheless, the number of erosion lesions seen in the spine was still insufficient, with the positive rate being only around one-thousandth. This indicated that the incidence of erosion lesions was low and there was a one-thousandth chance to lose 1 point due to there being no detection of erosion lesions occurring while using this mSASSS AI model, which was estimated to have little impact. It should be noted that the scoring of spinal erosion lesions was excluded in RASSSs and that spinal erosion lesions could possibly be negligible in real-world practice. Both the increasing number of labeled data and the refinement of annotation accuracy play an important role. Along with the application and promotion of the AI mSASSS prediction model, more labeled data can be collected, making it useful to train more AI models in the future.

Koo et al.[17] recently proposed a deep learning-based model for mSASSS assessment in AS patients in Korea, with good accuracy resulting in one corner of the vertebral body. However, their study did not differentiate each structural lesion with a score of 1 point in their AI model; therefore, a score of 1 may represent erosions as well as squaring or sclerosis. We distinguished each type of different structural damage and determined the total accuracy with satisfactory results. Additionally, cases of severe deformities or artificial structures were excluded in the Korea cohort, and the developed algorithm was not constructed using software, thus causing difficulty in clinical validation in medical settings. We scored all the structural lesions detected in the radiographs, and the results were more in line with real-world medical conditions. Furthermore, we will include the AI model in the application program, so there may be an opportunity to carry out medical verification in the future. It should be a top priority to use machine deep learning and AI to develop the best mSASSS prediction model for patients with AS, and to also automatically report mSASSSs in medical settings in order to better reduce interpretation errors and improve diagnostic accuracy. Previous studies have revealed that sclerosis change in sacroiliac joints correlated positively with lower back pain, stiffness, and sleep disturbance ($r = 0.45$, $p < 0.05$).[27] Through automatic quantification, an mSASSS prediction model would assist physicians in making both clinical judgments and medical decisions in real time. By combining prompt mSASSSs, clinical conditions, AS disease activity, physical examinations, and laboratory results, we were able to predict treatment efficacy, disease prognosis, and comorbidity

Table 3. The interobserver agreement of labeling for spinal lesions in patients with AS in majority decision.

Lesions	Total number of agreements plus disagreements	Number of agreements	Interobserver agreement	Intraobserver agreement
Erosion	27,768	27,542	99.2	99.8
Sclerosis	27,768	25,677	92.5	96.3
Squaring	27,768	23,312	84.0	93.9
Syndesmophyte	27,768	25,187	90.7	97.1
Bony bridge	27,768	26,669	96.0	99.1
Osteophyte	27,768	25,380	91.4	97.5

AS, ankylosing spondylitis.

Table 4. The training results of the validation set in patients with AS.

Lesions	Score	AUC-ROC	Specificity	Sensitivity	Accuracy
Cervical					
Bony bridge	3	0.998	0.988	0.974	0.985
Syndesmophyte	2	0.957	0.942	0.778	0.917
Erosion	1	NA	NA	NA	NA
Lumbar					
Bony bridge	3	0.994	0.983	0.971	0.974
Syndesmophyte	2	0.940	0.905	0.806	0.894
Erosion	1	NA	NA	NA	NA
Cervical and lumber					
Sclerosis	1	0.956	0.939	0.798	0.916
Squaring	1	0.944	0.922	0.776	0.886

AS, ankylosing spondylitis; AUC-ROC, area under the curve of the receiver operating characteristic curve; NA, not available.

Table 5. Confusion matrix on the testing set among two experienced rheumatologists, one experienced radiologist, and AI model.

Rheumatologists/ radiologists	AI model				Total	Accuracy
	0	1	2	3		
0	3496	269	144	38	3947	0.886
1	82	276	22	28	408	0.677
2	26	13	138	10	187	0.738
3	18	9	26	469	522	0.899
Total	3622	567	330	545	5064	0.865

AI, artificial intelligence.

occurrence, while also providing an individualized optimal treatment plan and integrated health care for each patient with AS.

Evidence has shown that there is low inter- and intra-observer reproducibility for scoring the radiographs of spine and sacroiliac joints in patients with AS.[27,28] Well-trained and experienced rheumatologists and radiologists were usually needed for mSASSS assessments, but the reliability of the results remained doubtful. Additionally, minimal structural changes seen on radiographs were initially difficult to detect with the human eye, so it required a long observation period after overwhelming radiographic damage had developed. Therefore, a trained and attentive AI system is necessary for mSASSS evaluation as a clinical aide. To improve accuracy, we have also attempted using the most advanced EfficientNet V2 for AI training. However, the accuracy was not as high as expected. Since the use of AI technology is currently booming, perhaps an ideal AI model that is particularly in line with mSASSS assessment will be released in the future.

The present study has demonstrated the use of the deep learning model for scoring mSASSS to detect the vertebral structural change in patients with AS through one label architecture, which could be put into clinical application in the future. Notably, spine X-rays from a normal or control population without AS were not included in our AI model development, and it cannot be applied to patients without AS.

There were some limitations in the study. First, although structural damage of the vertebrae seen on radiographs was interpreted by physicians across different hospitals, participants in the study were recruited from only a single center. External validation is still required and is now in progress at the Rheumatology and Immunology Center of China Medical University Hospital in Taiwan. Second, there was a shortage of certain structural lesion data in the mSASSS prediction model. However, we introduced a one label architecture AI model to enrich lesion size, allowing us to examine one single lesion at a time, which increased the accuracy remarkably. Third, there still remains certain difficulties that must be overcome regarding clinical implementation. Future plans include extending the study by incorporating a larger and more diverse patient population and exploring the model's applicability to sacroiliac joint radiographs. We will also focus on refining our AI model, including external validation, to address current limitations. Moreover, the inclusion of medical records, patient-reported outcomes, and environments such as lifestyles and genetic information could also be considered in future research. More studies are still necessary in order to explore any additional issues that may be related to clinical implementation of the mSASSS prediction model.

Conclusion

This study developed a well-trained deep learning model involving mSASSS scoring in order to detect the vertebral structural damage in patients with AS, with solid accuracy. The model would be further validated externally in the future for improving reliability in outcomes. An AI model for mSASSS scoring could provide accurate and real-time mSASSS evaluation, which is essential for advancing research and understanding the progression of AS. This tool can aid researchers in obtaining consistent and objective data, thereby enhancing the reproducibility and reliability of clinical studies.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Taichung Veterans General Hospital, Taiwan (IRB No. CE20295A) and waived requirement for informed consent because patient data were anonymized before analysis.

Consent for publication

Not applicable.

Author contributions

Yen-Ju Chen: Investigation; Methodology; Validation; Writing – original draft; Writing – review & editing.

Der-Yuan Chen: Conceptualization; Data curation; Investigation; Methodology; Project administration; Supervision; Writing – review & editing.

Haw-Chang Lan: Conceptualization; Data curation; Methodology; Project administration; Supervision; Writing – review & editing.

An-Chih Huang: Data curation; Software.

Yi-Hsing Chen: Writing – review & editing.

Wen-Nan Huang: Writing – review & editing.

Hsin-Hua Chen: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing – review & editing.

Acknowledgements

We would like to thank the Biostatistics Task Force of Taichung Veterans General Hospital for their assistance in performing the statistical analysis. The authors also sincerely appreciate the assistance we received from the Center for Translational Medicine of Taichung Veterans General Hospital.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by a grant from Taichung Veterans General Hospital, Taiwan (TCVGH-1113001C).

Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

The data in this study are available from the corresponding author under reasonable request.

ORCID iDs

Yen-Ju Chen https://orcid.org/0000-0002-9450-9144

Der-Yuan Chen https://orcid.org/0000-0003-1266-1423

An-Chih Huang https://orcid.org/0009-0003-6266-2368

Hsin-Hua Chen https://orcid.org/0000-0002-7304-4587

References

1. Navarro-Compan V, Sepriano A, El-Zorkany B, et al. Axial spondyloarthritis. *Ann Rheum Dis* 2021; 80: 1511–1521.
2. Taurog JD, Chhabra A and Colbert RA. Ankylosing spondylitis and axial spondyloarthritis. *N Engl J Med* 2016; 374: 2563–2574.
3. Landewe R, Dougados M, Mielants H, et al. Physical function in ankylosing spondylitis is independently determined by both disease activity and radiographic damage of the spine. *Ann Rheum Dis* 2009; 68: 863–867.
4. Maksymowych WP, Claudepierre P, de Hooge M, et al. Structural changes in the sacroiliac joint on MRI and relationship to ASDAS inactive disease in axial spondyloarthritis: a 2-year study comparing treatment with etanercept in EMBARK to a contemporary control cohort in DESIR. *Arthritis Res Ther* 2021; 23: 43.
5. Creemers MC, Franssen MJ, van't Hof MA, et al. Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. *Ann Rheum Dis* 2005; 64: 127–129.
6. Bakan BM, Sivas F, Inal EE, et al. Comparison of the bath ankylosing spondylitis radiology index and the modified stoke ankylosing spondylitis spine score in Turkish patients with ankylosing spondylitis. *Clin Rheumatol* 2010; 29: 65–70.
7. Baraliakos X, Listing J, Rudwaleit M, et al. Development of a radiographic scoring tool for ankylosing spondylitis only based on bone formation: addition of the thoracic spine improves sensitivity to change. *Arthritis Rheum* 2009; 61: 764–771.
8. Ramiro S, Claudepierre P, Sepriano A, et al. Which scoring method depicts spinal radiographic damage in early axial spondyloarthritis best? Five-year results from the DESIR cohort. *Rheumatology (Oxford)* 2018; 57: 1991–2000.
9. Ramiro S, van Tubergen A, Stolwijk C, et al. Scoring radiographic progression in ankylosing spondylitis: should we use the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) or the Radiographic Ankylosing Spondylitis Spinal Score (RASSS)? *Arthritis Res Ther* 2013; 15: R14.
10. van der Heijde D, Braun J, Deodhar A, et al. Modified stoke ankylosing spondylitis spinal score as an outcome measure to assess the impact of treatment on structural progression in ankylosing spondylitis. *Rheumatology (Oxford)* 2019; 58: 388–400.
11. Sepriano A, Regel A, van der Heijde D, et al. Efficacy and safety of biological and targeted-synthetic DMARDs: a systematic literature review informing the 2016 update of the ASAS/EULAR recommendations for the management of axial spondyloarthritis. *RMD Open* 2017; 3: e000396.
12. van der Heijde D and Landewé R. Selection of a method for scoring radiographs for ankylosing spondylitis clinical trials, by the Assessment in Ankylosing Spondylitis Working Group and OMERACT. *J Rheumatol* 2005; 32: 2048–2049.
13. Wanders AJ, Landewé RB, Spoorenberg A, et al. What is the most appropriate radiologic scoring method for ankylosing spondylitis? A comparison of the available methods based on the Outcome Measures in Rheumatology Clinical Trials filter. *Arthritis Rheum* 2004; 50: 2622–2632.
14. van der Heijde D, Baraliakos X, Hermann KA, et al. Limited radiographic progression and sustained reductions in MRI inflammation in patients with axial spondyloarthritis: 4-year imaging outcomes from the RAPID-axSpA phase III randomised trial. *Ann Rheum Dis* 2018; 77: 699–705.
15. Braun J, Baraliakos X, Deodhar A, et al.; MEASURE 1 Study Group. Effect of secukinumab on clinical and radiographic outcomes in ankylosing spondylitis: 2-year results from the randomised phase III MEASURE 1 study. *Ann Rheum Dis* 2017; 76: 1070–1077.
16. van der Heijde D, Østergaard M, Reveille JD, et al. Spinal radiographic progression and predictors of progression in patients with radiographic axial spondyloarthritis receiving ixekizumab over 2 years. *J Rheumatol* 2022; 49: 265–273.

17. Koo BS, Lee JJ, Jung JW, 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* 2022; 14: 1759720X221114097.

18. van der Linden S, Valkenburg HA and Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum* 1984; 27: 361–368.

19. Rudwaleit M, Landewé R, van der Heijde D, et al. The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part I): classification of paper patients by expert opinion including uncertainty appraisal. *Ann Rheum Dis* 2009; 68: 770–776.

20. Rudwaleit M, van der Heijde D, Landewé R, et al. The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. *Ann Rheum Dis* 2009; 68: 777–783.

21. Braun J and van der Heijde D. Imaging and scoring in ankylosing spondylitis. *Best Pract Res Clin Rheumatol* 2002; 16: 573–604.

22. Clark RA, Mentiplay BF, Hough E, et al. Three-dimensional cameras and skeleton pose tracking for physical function assessment: a review of uses, validity, current developments and Kinect alternatives. *Gait Posture* 2019; 68: 193–200.

23. Luvizon DC, Picard D and Tabia H. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE Trans Pattern Anal Mach Intell* 2021; 43: 2752–2764.

24. Urinbayev K, Orazbek Y, Nurambek Y, et al. End-to-end deep diagnosis of X-ray images. *Annu Int Conf IEEE Eng Med Biol Soc* 2020; 2020: 2182–2185.

25. Liang X, Peng C, Qiu B, et al. Dense networks with relative location awareness for thorax disease identification. *Med Phys* 2019; 46: 2064–2073.

26. von Elm E, Altman DG, Egger M, et al.; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 2007; 370: 1453–1457.

27. Taylor HG, Wardle T, Beswick EJ, et al. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. *Br J Rheumatol* 1991; 30: 330–335.

28. Hollingsworth PN, Cheah PS, Dawkins RL, et al. Observer variation in grading sacroiliac radiographs in HLA-B27 positive individuals. *J Rheumatol* 1983; 10: 247–254.