

## Deep Learning Based Automatic Ankle Tenosynovitis Quantification from MRI in Patients with Psoriatic Arthritis

### A Feasibility Study

Arbabi, Saeed; Arbabi, Vahid; Costa, Lorenzo; Katen, Iris ten; Mastbergen, Simon C.; Seevinck, Peter R.; de Jong, Pim A.; Weinans, Harrie; Jansen, Mylène P.; Foppen, Wouter

**DOI**

[10.3390/diagnostics15121469](https://doi.org/10.3390/diagnostics15121469)

**Publication date**

2025

**Document Version**

Final published version

**Published in**

Diagnostics

### Citation (APA)

Arbabi, S., Arbabi, V., Costa, L., Katen, I. T., Mastbergen, S. C., Seevinck, P. R., de Jong, P. A., Weinans, H., Jansen, M. P., & Foppen, W. (2025). Deep Learning Based Automatic Ankle Tenosynovitis Quantification from MRI in Patients with Psoriatic Arthritis: A Feasibility Study. *Diagnostics*, 15(12), Article 1469. <https://doi.org/10.3390/diagnostics15121469>

### Important note

To cite this publication, please use the final published version (if applicable).  
Please check the document version above.

### Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

### Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.  
We will remove access to the work immediately and investigate your claim.

## Article

# Deep Learning Based Automatic Ankle Tenosynovitis Quantification from MRI in Patients with Psoriatic Arthritis: A Feasibility Study

Saeed Arbabi <sup>1,2,\*</sup>, Vahid Arbabi <sup>2,3</sup> , Lorenzo Costa <sup>1,2</sup> , Iris ten Katen <sup>4</sup>, Simon C. Mastbergen <sup>5</sup> , Peter R. Seevinck <sup>1,6</sup>, Pim A. de Jong <sup>4,\*</sup> , Harrie Weinans <sup>2,7</sup> , Mylène P. Jansen <sup>5,†</sup>  and Wouter Foppen <sup>4,†</sup> 

<sup>1</sup> Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands; lcosta-2@umcutrecht.nl (L.C.); p.seevinck@umcutrecht.nl (P.R.S.)

<sup>2</sup> Department of Orthopedics, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands; v.arbabi@umcutrecht.nl (V.A.); h.h.weinans@umcutrecht.nl (H.W.)

<sup>3</sup> Orthopaedic-Biomechanics Research Group, Department of Mechanical Engineering, Faculty of Engineering, Birjand 561, Iran

<sup>4</sup> Department of Radiology, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands

<sup>5</sup> Department of Rheumatology & Clinical Immunology, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands; s.mastbergen@umcutrecht.nl (S.C.M.); m.p.jansen-36@umcutrecht.nl (M.P.J.)

<sup>6</sup> MRGuidance B.V., 3584 CX Utrecht, The Netherlands

<sup>7</sup> Department of Biomechanical Engineering, Delft University of Technology (TU Delft), 2628 CD Delft, The Netherlands

\* Correspondence: s.arbabi@umcutrecht.nl (S.A.); p.dejong-8@umcutrecht.nl (P.A.d.J.)

† These authors contributed equally to this work.

**Abstract: Background/Objectives:** Tenosynovitis is a common feature of psoriatic arthritis (PsA) and is typically assessed using semi-quantitative magnetic resonance imaging (MRI) scoring. However, visual scoring is variable. This study evaluates a fully automated, deep-learning approach for ankle tenosynovitis segmentation and volume-based quantification from MRI in psoriatic arthritis (PsA) patients. **Methods:** We analyzed 364 ankle 3T MRI scans from 71 PsA patients. Four tenosynovitis pathologies were manually scored and used to create ground truth segmentations through a human-machine workflow. For each pathology, 30 annotated scans were used to train a deep-learning segmentation model based on the nnUNet framework, and 20 scans were used for testing, ensuring patient-level disjoint sets. Model performance was evaluated using Dice scores. Volumetric pathology measurements from test scans were compared to radiologist scores using Spearman correlation. Additionally, 218 serial MRI pairs were assessed to analyze the relationship between changes in pathology volume and changes in visual scores. **Results:** The segmentation model achieved promising performance on the test set, with mean Dice scores ranging from 0.84 to 0.92. Pathology volumes correlated with visual scores across all test MRIs (Spearman  $\rho = 0.52$ – $0.62$ ). Volume-based quantification captured changes in inflammation over time and identified subtle progression not reflected in semi-quantitative scores. **Conclusions:** Our automated segmentation tool enables fast and accurate quantification of ankle tenosynovitis in PsA patients. It may enhance sensitivity to disease progression and complement visual scoring through continuous, volume-based metrics.

**Keywords:** tenosynovitis; MRI; deep learning



Academic Editor: Zhuhuan Zhou

Received: 27 April 2025

Revised: 26 May 2025

Accepted: 29 May 2025

Published: 9 June 2025

**Citation:** Arbabi, S.; Arbabi, V.; Costa, L.; Katen, I.; Mastbergen, S.C.; Seevinck, P.R.; de Jong, P.A.; Weinans, H.; Jansen, M.P.; Foppen, W. Deep Learning Based Automatic Ankle Tenosynovitis Quantification from MRI in Patients with Psoriatic Arthritis: A Feasibility Study. *Diagnostics* **2025**, *15*, 1469. <https://doi.org/10.3390/diagnostics15121469>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Tenosynovitis, characterized by the inflammation of the tendon's synovial sheath, is a pivotal aspect in psoriatic arthritis (PsA) [1]. Its early detection and monitoring are important for initiating timely treatment and preventing joint damage, thereby improving the patient's quality of life [2–5].

Imaging evaluation of tenosynovitis has advantages over clinical examinations of PsA and can reflect prodromal phase of PsA [1,6]. The Outcome Measures in Rheumatology (OMERACT) group developed the PsA MRI Scoring System (PsAMRIS), which includes tenosynovitis among other features and uses a semi-quantitative scoring scheme [7,8]. Despite the establishment of PsAMRIS, its discrete nature complicates careful monitoring of often subtle changes in disease activity [9,10]. Moreover, manually interpreting images is time-intensive and relies on the individual judgment of the radiologist [3,11–13]. A study by Glinatsi et al. validating the OMERACT PsAMRIS for the hand and foot in a randomized, placebo-controlled trial reported poor baseline interreader reliability for tenosynovitis in the foot (Intraclass Correlation Coefficient (ICC) 0.25–0.44), while good reliability was observed for other PsAMRIS features (ICC 0.72–0.96) [8]. Jin et al. demonstrate how automated segmentation can improve consistency [14].

Quantifying inflammatory features using volumetric and signal intensity-based metrics may offer a more continuous, objective alternative to discrete visual scoring. This approach could be more sensitive to variations when assessing disease activity and response to therapy over time.

In this study, we assess the feasibility of a fully automatic method for segmentation and quantification of MRI tenosynovitis in the foot. We hypothesize that a fully automated segmentation pipeline can serve as a reliable precursor to volume-based quantification of tenosynovitis, offering more sensitive monitoring than visual scoring.

Previous studies have attempted automatic segmentation and quantification of inflammatory pathologies on MRI images. Haj-Mirzaian et al. summarized the feasibility of computer-assisted quantification of MRI inflammatory arthritis pathologies as the responsiveness of scoring methods are limited by their discrete nature and that those systems can be insensitive to early inflammatory changes [3]. They expect that artificial intelligence (AI)-driven approaches, such as ones based on deep learning (DL) algorithms may overcome these limitations [15,16]. Moreover, Momtazmanesh et al. surveyed the use of AI in rheumatoid arthritis and found no studies using DL-based method for automatic segmentation of inflammatory pathologies [17,18]. Schlereth et al. used deep learning for the classification of erosion, synovitis and osteitis in hand MRI of patients with inflammatory arthritis [19]. Aizenberg et al. investigated the feasibility of automatic quantification of tenosynovitis on MRI of the wrist in patients with early arthritis. They used an atlas-based method for segmentation of tendons, extracted the ROIs around tendons and applied fuzzy clustering to identify voxel intensities inside ROI that can be associated with inflammation [20–22]. More recently, Hepburn et al. introduced a human-machine workflow for semi-automatic segmentation and quantification of inflammation load in spondyloarthritis [11]. To the best of our knowledge, no previous studies evaluated the feasibility of DL-based, fully automatic segmenting of tenosynovitis from ankle MRI in inflammatory arthritis.

In our study, we assess the feasibility of a fully automatic method for segmentation and quantification of tenosynovitis on MRI. This method relies on automatic, DL-based segmentation of tendons in the ankle instead of being reliant on manual expert pathology segmentation for ground truth generation, which is a very time-consuming task.

## 2. Materials and Methods

This is a retrospective study to determine the feasibility of automatic segmentation and quantification of ankle tenosynovitis. The quantification is done by automatic segmentation of tenosynovitis and its regions of interest.

### 2.1. Data

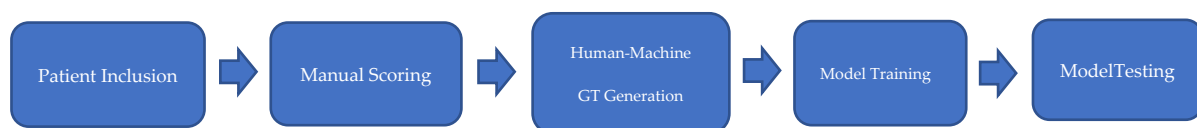
Serial bilateral (left and right separately) ankle MRI data from 71 patients with PsA, collected across multiple centers as part of the TOFA-PREDICT study (EudraCT Number 2017-003900-28), were utilized. All the patients included in this study provided written consent and the study was approved by the Medical Research Ethics Committee in Utrecht, Netherlands (MREC reference number: NL63439.041.17). Despite being multi-center, the used images had been acquired from the same MRI machine manufacturer with approximately the same parameters. We note that most images were acquired using Philips Ingenia 3T scanners (Philips Healthcare, Best, The Netherlands) across different sites. One scan acquired with a Siemens 1.5T scanner was excluded from analysis to maintain consistency in field strength and acquisition protocol. The specifics of the study have been described in prior publication [23].

A total of 364 3T T1-weighted Proton Density Spectral Adiabatic Inversion Recovery (PD SPAIR) ankle MR images (coming from the 71 patients, each with multiple timepoints and bilateral images) were analyzed for scoring and quantification purposes. Details on the MRI parameters are reported in Table S1 of the Supplementary Material.

Quantification of tenosynovitis around 4 tendon regions, namely the tibialis posterior tendon, flexor digitorum longus tendon, flexor hallucis longus tendon, peroneal longus/brevis tendon, was the topic of this study.

Two musculoskeletal radiologists (WF, IK) conducted scoring for tenosynovitis around the abovementioned four specific tendon regions, following the guidelines of PsAMRIS, which were adapted by the study team for assessment of tenosynovitis around the ankle joint [7,8]. Tenosynovitis was assessed using semiquantitative assessment of fluid within the tendon sheaths (0: none, 1:  $<1/2$  tendon thickness, 2:  $\geq 1/2$  tendon thickness, 3:  $\geq 1$  tendon thickness). In cases of disagreement regarding the tenosynovitis scores, the differences were resolved through follow-up consensus readings.

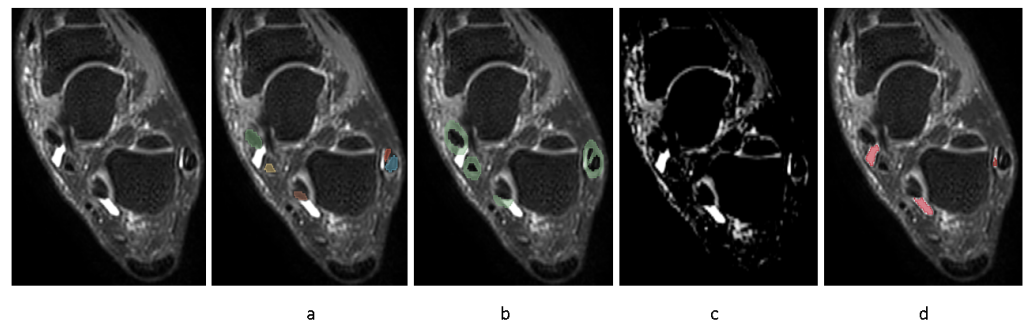
Diagram of the workflow used in this study is as shown in Figure 1.



**Figure 1.** Workflow of the study.

### 2.2. Ground Truth (GT)

In order to automate the segmentation of tenosynovitis pathologies and validating this process, a dataset of ground truth tenosynovitis segmentation was required. Tenosynovitis ground truth segmentation as shown in Figure 2 was performed in a human-machine workflow to speed up the generation of dataset and to increase the consistency across readers. In the human-machine workflow, a radiologist (WF) supervised an image processing pipeline that generates initial pathology segmentations. These segmentations were then corrected by an MD researcher with 5 years of experience in orthopedics imaging (LC) and approved by radiologist (WF).



**Figure 2.** Human-machine workflow of creating ground truth pathology segmentation dataset. (a–d): (a): tendons segmented (colors), (b): ROIs around tendons defined (green), (c): clustering highlights regions (lighter regions), (d): highlighted regions inside ROIs selected (red).

The details of the human-machine workflow are as follows: In the first step, as foot tenosynovitis pathologies are around any of 4 tendon regions, these 4 structures were segmented. As peroneal tenosynovitis occurs around both peroneal longus and peroneal brevis tendons, both these structures were segmented. This step is shown in Figure 2a. An MD researcher with 5 years of experience in orthopedics imaging (LC) performed manual segmentations of tendons on a set of 50 MRIs utilizing the 3D Slicer software, version 5.0.1. All 50 tendon segmentations were then reviewed by one musculoskeletal radiologist (WF).

In the majority of cases, the initial segmentations generated by the pipeline were visually plausible and required only minor boundary adjustments or removal of false positives. In fewer cases (~20%), more substantial edits were needed to better capture the extent of inflammation. The human-machine workflow significantly reduced annotation time while ensuring consistency through expert oversight.

Those 50 MRIs (from different patients) were used for segmentation training/testing, while the remaining scans were used for correlation and longitudinal analysis. In order to automate this first step, an nnU-Net [24,25] was trained on 30 full segmentations and tested on 20 full segmentations. nnUNet is a segmentation framework based on U-Net framework [26] that automatically configures the hyperparameters based on the fingerprint of the dataset. The segmentation of tendons is helpful as well in quantification of tenosynovitis, as according to PsAMRIS, tenosynovitis is scored based on inflammation thickness proportionate to tendon thickness.

In step 2, ROIs around the segmented tendons were automatically extracted, as depicted in Figure 2b.

In step 3, fuzzy c-means clustering was applied to the image voxels with the assumption of two clusters (brighter/darker voxels). Then, voxel intensities surpassing the higher cluster center value inside ROI underwent a threshold optimization for probability of those voxels belonging to the cluster. The illustrative result of this clustering method is encapsulated in Figure 2c. The threshold values that yielded the highest correlation of pathology volume with the radiologists' semi-quantitative pathology scores were taken for further consideration by musculoskeletal radiologist (WF). The radiologist reviewed potential threshold levels on a selection of 10 MRIs from training set, representing various degrees of severity, to identify a probability threshold that most accurately captures the presence of pathology. We applied the same threshold across all four tendon regions, as the signal characteristics of tenosynovitis in these ROIs were visually consistent across images acquired with the same MRI protocol. Importantly, all resulting segmentations were reviewed and, if needed, corrected by an expert reader, ensuring validity regardless of the initial threshold. This ensured that the final annotations used for model training were consistent and accurate. While more advanced, adaptive thresholding techniques may

improve automation further, our fixed-threshold approach provided sufficient accuracy for the purposes of this feasibility study.

In step 4, using the chosen threshold value in previous step, inflammatory voxels were automatically segmented and retained if situated within the disease-specific ROI from step 2 or emanated from within this defined area. This is shown in Figure 2d.

Finally, segmentations resulting from step 4 underwent a thorough review of 50 images and any necessary adjustments by MD researcher with 5 years of experience in orthopedics imaging (LC), instructed by a musculoskeletal radiologist (WF). The segmentations were validated and approved by the radiologist (WF).

### 2.3. Data Portions

At the end of the workflow, for each tenosynovitis pathology, a dataset of 50 expert-approved segmentation was thus ready for automation. Thirty segmentations were used for training the segmentation model, and twenty were used for testing the model. To evaluate the generalizability of our segmentation model, we employed a 2-fold cross-validation approach. Specifically, our dataset of 50 labeled segmentations was divided into two folds. For each fold, we used 30 segmentations for training and 20 segmentations for testing. This ensured that all segmentations were used in both training and testing at least once. Each segmentation was included in the testing set once at most, allowing us to obtain a mean test result and assess the model's performance across different subsets of the data. Data in training and test sets were from different patients to avoid information leak.

### 2.4. Model Training

NnU-Net is inherently structured to compensate for the absence of testing data, employing a 5-fold cross-validation approach as its standard protocol [27]. Given that we had access to an independent testing set within the same dataset, we bypassed this default feature and conducted training using the entirety of the training data. We allocated a distinct dataset for each pathology, processing each one separately. The semi-automatically corrected segmentations were employed to train a deep learning nnU-Net network with 3D full resolution configuration for 1000 epochs using a Dice similarity coefficient with Cross Entropy loss. To enhance the network's resilience to the various MRI artifacts present in the dataset, the TorchIO data augmentation library [28] was utilized to add noise, ghosting effect or bias field inhomogeneity to images and create augmented image-segmentation pairs. The nnU-Net framework automatically integrates optimal pre-processing and architectural strategies by analyzing characteristics of the given dataset, such as modality, spacing and dimensions of the scans [29].

We selected nnU-Net as our segmentation framework due to its demonstrated state-of-the-art performance across diverse medical image segmentation tasks and its automatic configuration of preprocessing, architecture and training parameters. In our feasibility context with a limited dataset, nnU-Net provided a robust and reproducible pipeline without the need for extensive manual tuning. Although alternative models such as attention-based U-Nets or transformer architectures may offer improvements in specific cases, these were not explored here due to the scope and focus of the study. A recent comparative analysis found that nnU-Net outperformed attention U-Nets and Res-U-Nets across tasks like tumor and polyp segmentation, particularly in recall and Dice score [30].

The code and models used in the study are available at <https://github.com/sarbabi/TenosSeg> accessed 1 April 2025.

### 2.5. Experiments

Evaluation of automated against semi-automated pathology segmentation.



To assess the efficacy of the trained nnU-Net models, we used them to predict the segmentation of the test scans and compared them with the semi-automatically generated ground truth segmentations available. The training and test sets were selected in a way that covers different scores of tenosynovitis. The test set did not have any patients in common with the training set to avoid information leak. For quantitative evaluation of performance, we employed the Dice coefficient. This indicates how well the trained model can segment the pathology compared to the expert.

### *2.6. Evaluation of Automated Tendon Segmentation*

The performance of automated segmentation of tendon regions directly affects the segmentation of ROIs. So, in this part, we evaluate the tendon segmentation model's performance. The training and tests set were selected randomly from images. The model's accuracy was benchmarked against the 20 manually segmented MRIs, employing the Dice similarity coefficient as a metric for comparison.

Evaluating human-machine workflow compared to manual method for pathology ground truth segmentations.

To evaluate the consistency and accuracy of the radiologist-instructed pathology segmentations using human-machine workflow, a comparison was made between fully manual segmentations and human-machine workflow. A radiologist (WF) and a radiologist-instructed MD researcher (LC) each performed manual segmentations of tibialis posterior tenosynovitis on 10 randomly selected images from different severities of disease on two separate occasions, one month apart to minimize recall bias. In addition, they each conducted a single segmentation session where they refined machine-generated initial pathology segmentations. The aim of this setup was to assess whether the human-machine workflow might influence the consistency and accuracy of the segmentations as compared to the traditional manual approach. The intra- and inter-observer agreement were calculated using Dice similarity coefficient.

Evaluating correlations of pathology volumes with scores.

Spearman rank correlations between volumes extracted from automatic segmentation of pathologies and the visual scores by radiologists were calculated, to indicate the level of alignment between human qualitative scores and machine quantifications.

### *2.7. Comparing the Sensitivity of Quantifications to Change in Tenosynovitis Score*

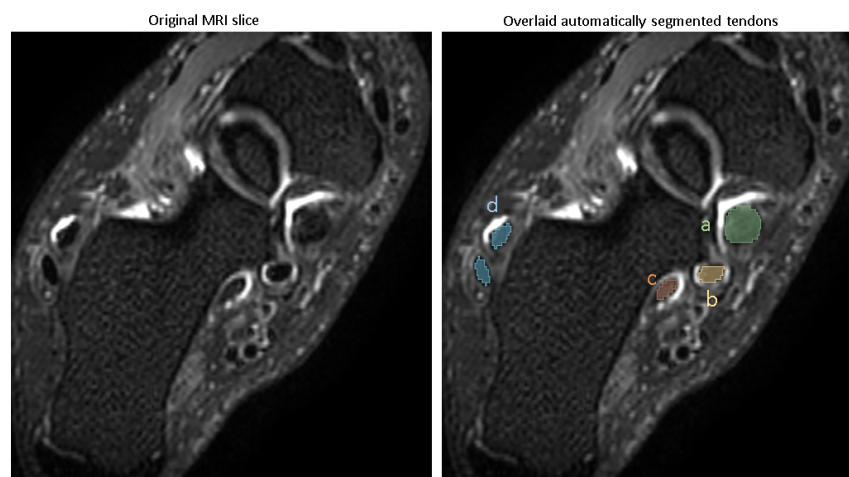
In order to provide insight into the sensitivity of our methods to detect changes over time in comparison to semi-quantitative scores, the change of inflammation score and volume were illustrated from baseline to week 16, and from week 16 to week 52, in patients where the radiologists unanimously agree on the baseline score. The disease progress, in terms of volume and score on 218 pairs of serial MRIs in 71, patients was shown.

## **3. Results**

The MRI parameters for the used PD SPAIR sequence of different centers are reported in Table S1 of Supplementary Materials. A total of 71 patients evaluated using 3T MRI (all centers except five) were included in this study. All included patients were evaluated on MRI scanners from the same manufacturer and using approximately the same parameters for the PD SPAIR sequence. The distribution of scores of radiologists is as depicted in Table S2 of Supplementary Materials.

Performance of the tendon segmentation model.

A representative segmentation of four tendon regions is demonstrated in Figure 3.



**Figure 3.** Automatic segmentation of tendons. a: Tibialis posterior tendon, b: flexor digitorum tendon, c: flexor hallucis longus tendon, d: peroneal longus/brevis tendon.

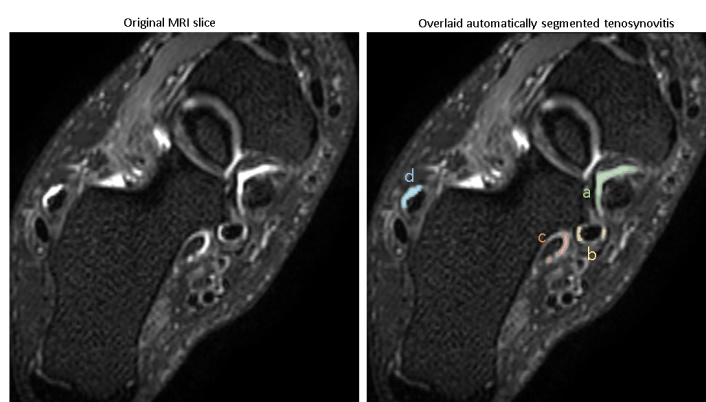
Evaluating the model's performance with the Dice coefficient on a test set of 20 images disjointed from the training set on the patient level (Table 1), showed that the model achieved a Dice score of (mean  $\pm$  SD) ( $0.94 \pm 0.01$ ) for the tibialis posterior tendon, ( $0.93 \pm 0.02$ ) for the flexor digitorum longus, ( $0.91 \pm 0.02$ ) for the flexor hallucis longus tendon and ( $0.95 \pm 0.01$ ) for the peroneal longus/brevis tendon, reflecting excellent segmentation accuracy.

**Table 1.** Dice coefficients of automatic segmentation of tendons.

Tendon	Dice Score
Tibialis posterior	$0.94 \pm 0.01$
Flexor digitorum longus	$0.93 \pm 0.02$
Flexor hallucis longus	$0.91 \pm 0.02$
Peroneal longus/brevis	$0.95 \pm 0.01$

### 3.1. Automating Pathology Segmentation

The visual examples of the results of the automatic pathology segmentations for an image are depicted in Figure 4.



**Figure 4.** Example image overlaid with automatic segmentation of 4 tenosynovitis pathologies. (a) Tibialis posterior tenosynovitis (score: 3), (b) flexor digitorum tenosynovitis (score: 1), (c) flexor hallucis longus tenosynovitis (score: 1), (d) peroneal longus/brevis tenosynovitis (score: 1).



The trained models were used for predicting pathologies in the test set, and the Dice scores reported the similarity of the predicted segmentation with the ground truth segmentation. The accuracy of segmentation in terms of Dice score (mean  $\pm$  SD) is shown in Table 2, with all pathologies showing mean Dice scores of  $>0.8$ .

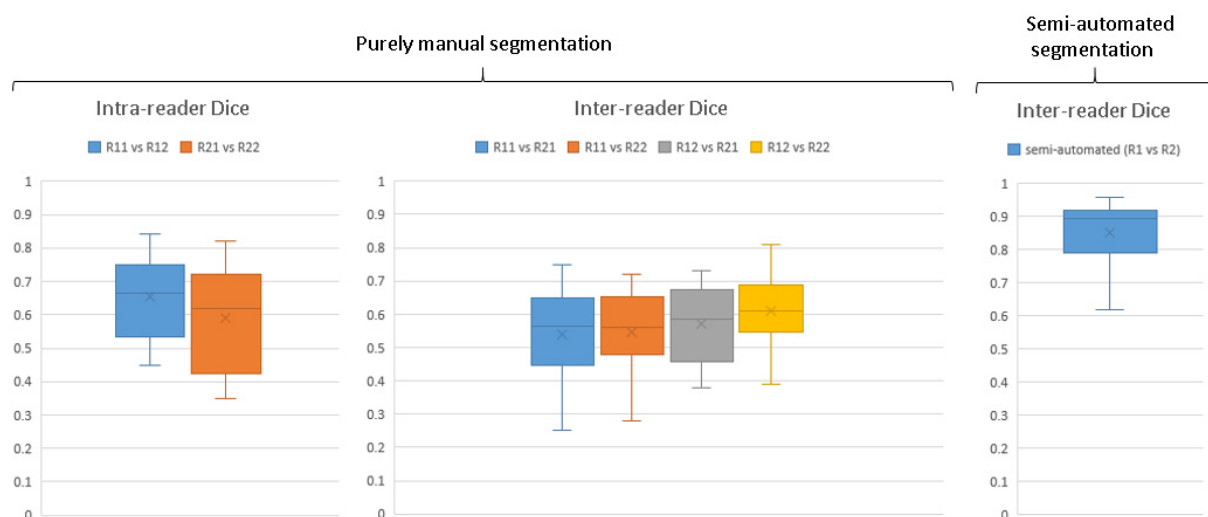
**Table 2.** Dice scores on test set for different pathologies (mean  $\pm$  SD).

Pathology	Test Dice
Tibialis posterior tenosynovitis	0.91 $\pm$ 0.04
Flexor digitorum tenosynovitis	0.84 $\pm$ 0.07
Flexor hallucis longus tenosynovitis	0.85 $\pm$ 0.06
Peroneal longus/brevis tenosynovitis	0.92 $\pm$ 0.04

These Dice scores indicate that the models are capable of reliably segmenting tenosynovitis pathologies in the ankle across different regions, with high overlap compared to expert-validated reference segmentations.

### 3.2. Evaluation of Human–Machine Workflow Compared with Manual Segmentation

The ground truth segmentations for automating tenosynovitis segmentation were generated using a human–machine workflow. Figure 5 illustrates the enhancement in ground truth pathology segmentation accuracy achieved with the aid of the human–machine workflow, quantified using Dice coefficients. For entirely manual segmentations carried out independently by two observers, the inter-reader agreement in terms of Dice scores varied widely, ranging from 0.25 to 0.81. The median Dice score for repeated segmentations by the same observer (intra-reader agreement) was recorded at 0.66 and 0.62, while the median for segmentations between different observers (inter-observer agreement) ranged from 0.56 to 0.61.



**Figure 5.** Fully manual and semi-automatic (human–machine) tibialis posterior tenosynovitis segmentation performance. Rij denotes reader i in j-th repetition.

Upon the implementation of the human–machine collaborative workflow, there was a noticeable improvement in the median inter-observer agreement Dice scores, rising to 0.89. This shows a significantly higher agreement between readers for correcting segmentations provided by algorithm than fully manual segmentation of pathology and thereby a need for less refinement by the radiologist in dataset generation.

### 3.3. Threshold Optimization Step

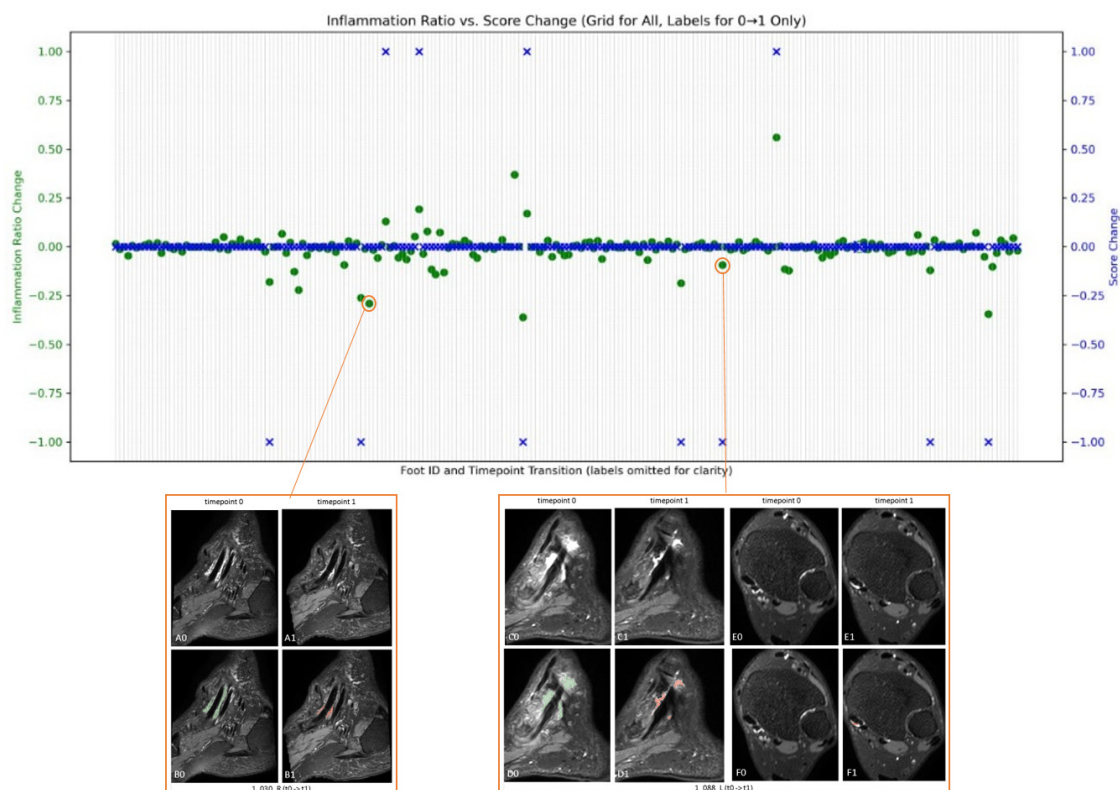
Figure S1 shows the results of the threshold optimization process for tibialis posterior tenosynovitis. As we see, with threshold values between around 0.67 and 0.84, the correlation between pathology volume and radiologist scores is stronger. The radiologist further reviewed this range of thresholds for 10 images stratified from different severities of pathology and based on a threshold value of 0.7 being chosen for tibialis posterior. The same threshold was used for segmenting tenosynovitis in all four regions.

### 3.4. Assessing Correlation Between Scores and Volumes

Evaluating Spearman rank correlation between scores and pathology volume in each ROI for all patients shows  $\rho = 0.62$  for tibialis posterior, 0.58 for flexor digitorum longus, 0.52 for flexor hallucis longus and 0.59 for peroneal longus/brevis, with  $p$ -values  $< 1 \times 10^{-16}$ , showing a moderate to strong correlation. While moderate, these correlations suggest that continuous volume-based measurements align meaningfully with expert semi-quantitative assessments and may provide a more sensitive alternative for longitudinal monitoring.

### 3.5. Assessing Inflammation Dynamics Through Volume Feature and Score Variations

The disease progress in terms of volume and score on 218 pairs of serial MRIs were evaluated for 71 patients. Figure 6 shows the changes in tibialis posterior tenosynovitis volume and tibialis posterior tenosynovitis scores over time.



**Figure 6.** The changes in volume along with scores in all patients. Blue crosses indicate score changes over time as scored by a radiologist, while green dots indicate volume changes as automatically segmented. This figure also illustrates one image that shows negative change in the volume of pathology between the two time-points while the score does not change. Also another example where score negatively changes, but the change in volume is not big.

For patient 1\_030\_R, depicted in the bottom left of Figure 6, although at both times the pathology is scored as 1 (score change is 0), the quantification suggests a decrease in

inflammation volume from baseline to week 16. In the same way, for patient 1\_088\_L, illustrated in the bottom right, the baseline score is 3, while at week 16, it is scored as 2. While the overlaid automatic pathology segmentation also depicts this change in C0 to C1, it also indicates a possible increase in pathology shown in E0 to E1.

In summary, the model achieved high segmentation performance across all target pathologies, with Dice scores exceeding 0.8. Volume measurements derived from these segmentations showed statistically significant moderate correlations with radiologist-assigned visual scores. In the longitudinal analysis, volume changes were observed in cases with and without changes in visual score, suggesting the potential sensitivity of continuous quantification for tracking disease activity. These findings support the feasibility of using deep learning-based volume quantification to complement conventional scoring methods.

#### 4. Discussion

This research assesses the feasibility of the application of AI in evaluation and monitoring of ankle tenosynovitis through MRI. The findings in this study indicate that the developed automated system could augment the traditional semi-quantitative scoring with more change-sensitive measurements of tenosynovitis, offering quicker results with reduced dependence on human judgment. This is consistent with existing research, which shows AI's potential to address the challenges of subjective assessments by reducing human error and bias in image analysis and offering continuous-scale measurements.

Although the automated segmentation models achieved Dice scores above 0.8 and moderate to strong correlations with visual scoring, these metrics do not imply perfect agreement or direct clinical interchangeability. The model-derived volume measures offer a continuous, objective view of inflammation burden that complements—but does not replace—expert assessment.

The implementation of a collaborative workflow integrating human expertise and machine capabilities has markedly improved the uniformity with which various observers pinpoint specific regions of interest in pathology images. This improvement implies that the synergy between human knowledge and machine generation of initial segmentations can substantially enhance the precision of data annotations. Our findings demonstrate that the human-machine workflow not only expedited dataset creation but also significantly improved consistency across readers. Compared to fully manual segmentation, which showed a wide range of inter-observer agreement (Dice: 0.56–0.61), the collaborative review of model-generated segmentations led to a much higher agreement (Dice: 0.89). We believe that the model's initial segmentation served as a stable reference, reducing subjective variability and ensuring more reproducible delineation of pathology—particularly in cases with subtle or borderline inflammation.

The automatic segmentation and volume quantification of tenosynovitis showed moderate to strong correlation with visual scores of tenosynovitis. This shows that our automatic quantification tool aligns favorably with human expert scores, while providing continuous data, which enables more accurate evaluation of changes over time. While Spearman correlations between visual scores and segmented pathology volumes ranged from 0.52 to 0.62, this moderate strength is consistent with expectations given the fundamental differences in data representation. Visual scores are discrete, semi-quantitative values with known inter-reader variability, whereas automated volume quantification provides continuous measures of inflammation. These differences inherently limit perfect correlation. However, the observed correlations—combined with longitudinal consistency—suggest that volume-based quantification may serve as a useful complementary tool to visual scoring, with potential to improve sensitivity in monitoring inflammation over time. Further validation is needed to determine its ability to detect subtle or early changes.

Furthermore, the AI model's ability to detect volume changes in cases where visual scores remain stable suggests potential sensitivity to subtle inflammatory dynamics. However, these volume fluctuations require further validation—such as radiologist re-scoring or clinical outcome studies—to confirm their clinical relevance. Such enhancements complement traditional radiological assessments, reducing the likelihood of overlooked cases. This is especially evident in the case studies where the quantification of inflammation showed changes over time that were not reflected in the tenosynovitis scores. However, the clinical relevance of the subtle abnormalities and changes over time as provided by quantifications needs further studies on minimally detectable changes and minimally clinically important difference.

Our automated platform is capable of performing quick tenosynovitis pathology segmentation and augmenting the scoring with segmentation for each image.

There are limitations to consider regarding this study. The main limitation is the small pool of expert-validated images for training and testing, attributed to the limited availability of expert validations. Future studies would benefit from a larger dataset. Despite this, the method demonstrated encouraging outcomes across 20 validation sets for each pathology and showed minimal signs of overfitting where there was no large unexpected difference between mean Dice scores in the testing and training sets. Differences in mean dice score between the training and testing sets were between 2.4% to 5.6% for different tenosynovitis pathologies. Another limitation arises from the fact that only little heterogeneity is present in our dataset. Although our data come from different sites, the MRIs are from the same MRI machine model and were acquired with very similar parameters. Only site 5, with one patient, had a different machine, which we did not include in analysis, since our inclusion criterion had been 3T MR images. It remains to be investigated how well the method performs on scans from other vendors. Further validation in independent cohorts and across scanner vendors is required to confirm generalizability for clinical deployment.

An additional challenge encountered in this study was the presence of MRI artifacts, including issues with fat suppression and the ghosting effect. However, assessment of fluid within the tendon sheaths, which appear as hyperintense in the fluid-sensitive sequences used, is not hampered by suboptimal fat suppression. In addition, we employed data augmentation strategies to accommodate the image discrepancies caused by these artifacts. The effects of the artifacts are therefore expected to be limited.

A further limitation stems from the two-stage image-scoring process, where radiologists first assess each MRI independently, followed by a consensus meeting in case of differing opinions. Although this method helps standardize evaluations, it might inadvertently lead to an artificially high agreement rate due to a learning effect at the outset. The act of collaborating to reach a consensus could produce a higher agreement rate than in scenarios where consensus is reached only after all MRIs have been reviewed. As a result, the inter-reader agreement rates reported might not accurately reflect the range of variability that might be seen in routine clinical practice or other research environments.

These limitations highlight the need for a meticulous interpretation of the study's results and identify areas for improvement in future research.

As this was a feasibility study, our work focused on implementing and validating a self-configuring segmentation pipeline using nnU-Net. While nnU-Net provides strong baseline performance with minimal manual tuning, future work could explore alternative architectures—such as attention-based U-Nets or transformer-based models—which may offer enhanced sensitivity to small or ambiguous pathological features. In addition, benchmarking against external segmentation frameworks and commercial or open-source tools would help assess generalizability and identify areas where our pipeline may be improved or adapted for broader clinical use.

Despite these limitations, our results suggest that deep learning-based quantification of tenosynovitis is feasible and potentially valuable as a complementary tool for objective inflammation tracking in research settings. Further validation in independent cohorts and expanded clinical studies is needed to confirm its robustness and applicability.

## 5. Conclusions

In conclusion, this study assessed the feasibility of augmenting tenosynovitis scoring with AI-generated pathology segmentation. The findings from this feasibility research highlight the potential of AI to automatically quantify ankle tenosynovitis on MRI with accurate pathology segmentation masks. The deep learning-based method showed moderate to strong agreement with visual scores and was able to track longitudinal changes in inflammation. While this is promising, further validation on independent datasets, exploration of clinical thresholds, and studies on reader reliability are needed before broader clinical adoption.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/diagnostics15121469/s1>, Table S1. Center inclusions, scanners, MRI parameters. Table S2. The number of images scored with each grade for each pathology. Figure S1. Threshold optimization for pathology extraction.

**Author Contributions:** Conceptualization, V.A., P.A.d.J., H.W., M.P.J. and W.F.; methodology, S.A., P.R.S., H.W., M.P.J., W.F.; validation, S.A., L.C., I.t.K., M.P.J., W.F.; formal analysis, S.A., M.P.J., W.F.; investigation, S.A., S.C.M., P.R.S., P.A.d.J., H.W., M.P.J., W.F.; resources, S.C.M., M.P.J.; data curation, S.A., L.C., M.P.J.; writing—original draft preparation, S.A., M.P.J., W.F.; writing—review and editing, V.A., L.C., I.t.K., S.C.M., P.R.S., P.A.d.J., H.W., M.P.J., W.F.; visualization, S.A., M.P.J., W.F.; supervision, P.R.S., P.A.d.J., H.W., M.P.J., W.F.; project administration, S.C.M., H.W., P.A.d.J.; funding acquisition, W.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was conducted as a collaboration between University Medical Center Utrecht and Pfizer. University Medical Center Utrecht is the study sponsor, and the project is co-funded by the PPP Allowance made available by Health-Holland Top Sector Life Sciences and Health, to stimulate public–private partnerships (grant number: LSHM17074).

**Institutional Review Board Statement:** This study was approved by the Medical Research Ethics Committee (MREC) in Utrecht, the Netherlands, under reference number NL63439.041.17. All patients included in the study provided written informed consent. Date: 2018-01-25.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patients to publish this paper.

**Data Availability Statement:** The MRI data used in this study were collected as part of the TOFA-PREDICT trial and contain sensitive patient information. Due to ethical and legal restrictions related to patient privacy and the conditions of the informed consent, the data are not publicly available. Access to the data may be requested from the corresponding author, subject to approval by the appropriate institutional review boards and data use agreements.

**Acknowledgments:** We would like to thank the TOFA-PREDICT study authors and author group: Harald E. Vonkeman (Medisch Spectrum Twente, Enschede; University of Twente, Enschede), Shasti C. Mooij (Medisch Spectrum Twente, Enschede), Lydia Schipper (Elisabeth- TweeSteden Hospital, Tilburg), Amin Herman (Antonius Hospital, Utrecht), Astrid M. van Tubergen (Maastricht University Medical Center, Maastricht; Maastricht University, Maastricht), Arno W. van Kuijk (Reade, Amsterdam; University of Amsterdam, Amsterdam), Tim L.T.A. Jansen (VieCuri MC, Venlo), Antoaneta C. Comarniceanu (VieCuri MC, Venlo), Sandra T.A. van Bijnen (Elisabeth- TweeSteden Hospital, Tilburg), Siska Wijngaarden (Gelre Hospitals, Apeldoorn), Simone A. Vreugdenhil (Antonius Hospital, Utrecht), Jacqueline S.L. Kloth (Diaconessenhuis, Utrecht), Emmerik F.A. Leijten (Maartenskliniek, Woerden), Z. Nazira Jahangier (Tergooi Hospital, Hilversum), Kavish J. Bhansing (St. Jansdal Hos-

pital, Harderwijk), Marc R. Kok (Maasstad Hospital, Rotterdam), Radjesh J. Bisoeendial (Maasstad Hospital, Rotterdam), Lenny Geurts-van Bon (ZGT, Almelo).

**Conflicts of Interest:** The company was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Zabotti, A.; Tinazzi, I.; Aydin, S.Z.; McGonagle, D. From Psoriasis to Psoriatic Arthritis: Insights from Imaging on the Transition to Psoriatic Arthritis and Implications for Arthritis Prevention. *Curr. Rheumatol. Rep.* **2020**, *22*, 24. [\[CrossRef\]](#)
2. Takata, T.; Takahashi, A.; Taniguchi, Y.; Terada, Y.; Sano, S. Detection of Asymptomatic Enthesitis in Psoriasis Patients: An Onset of Psoriatic Arthritis? *J. Dermatol.* **2016**, *43*, 650–654. [\[CrossRef\]](#)
3. Haj-Mirzaian, A.; Kubassova, O.; Boesen, M.; Carrino, J.; Bird, P. Computer-Assisted Image Analysis in Assessment of Peripheral Joint MRI in Inflammatory Arthritis: A Systematic Review and Meta-Analysis. *ACR Open Rheumatol.* **2022**, *4*, 721–734. [\[CrossRef\]](#)
4. Bagel, J.; Schwartzman, S. Enthesitis and Dactylitis in Psoriatic Disease: A Guide for Dermatologists. *Am. J. Clin. Dermatol.* **2018**, *19*, 839–852. [\[CrossRef\]](#)
5. Nicoara, A.-I.; Sas, L.-M.; Bită, C.E.; Dinescu, S.C.; Vreju, F.A. Implementation of Artificial Intelligence Models in Magnetic Resonance Imaging with Focus on Diagnosis of Rheumatoid Arthritis and Axial Spondyloarthritis: Narrative Review. *Front. Med.* **2023**, *10*, 1280266. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Hunt, L.; Eugénio, G.; Grainger, A.J. Magnetic Resonance Imaging in Individuals at Risk of Rheumatoid Arthritis. *Best Pract. Res. Clin. Rheumatol.* **2017**, *31*, 80–89. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Østergaard, M.; McQueen, F.; Wiell, C.; Bird, P.; Bøyesen, P.; Ejbjerg, B.; Peterfy, C.; Gandjbakhch, F.; Duer-Jensen, A.; Coates, L.; et al. The OMERACT Psoriatic Arthritis Magnetic Resonance Imaging Scoring System (PsAMRIS): Definitions of Key Pathologies, Suggested MRI Sequences, and Preliminary Scoring System for PsA Hands. *J. Rheumatol.* **2009**, *36*, 1816–1824. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Glinatsi, D.; Bird, P.; Gandjbakhch, F.; Mease, P.J.; Bøyesen, P.; Peterfy, C.G.; Conaghan, P.G.; Østergaard, M. Validation of the OMERACT Psoriatic Arthritis Magnetic Resonance Imaging Score (PsAMRIS) for the Hand and Foot in a Randomized Placebo-Controlled Trial. *J. Rheumatol.* **2015**, *42*, 2473–2479. [\[CrossRef\]](#)
9. Kleinrensink, N.J.; Foppen, W.; Katen, I.T.; Leijten, E.F.A.; de Jong, P.A.; Spierings, J. Longitudinal Follow-Up Using the Heel Enthesitis Magnetic Resonance Imaging Scoring System (HEMRIS) Shows Minimal Changes in Heel Enthesitis Assessed in Spondyloarthritis and Psoriasis Patients. *J. Pers. Med.* **2022**, *12*, 1765. [\[CrossRef\]](#)
10. Mathew, A.J.; Østergaard, M. Magnetic Resonance Imaging of Enthesitis in Spondyloarthritis, Including Psoriatic Arthritis—Status and Recent Advances. *Front. Med.* **2020**, *7*, 296. [\[CrossRef\]](#)
11. Hepburn, C.; Jones, A.; Bainbridge, A.; Ciurtin, C.; Iglesias, J.E.; Zhang, H.; Hall-Craggs, M.A.; Bray, T.J.P. Volume of Hyperintense Inflammation (VHI): A Quantitative Imaging Biomarker of Inflammation Load in Spondyloarthritis, Enabled by Human-Machine Cooperation. *PLoS ONE* **2023**, *18*, e0284508. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Zheng, Y.; Bai, C.; Zhang, K.; Han, Q.; Guan, Q.; Liu, Y.; Zheng, Z.; Xia, Y.; Zhu, P. Deep-Learning Based Quantification Model for Hip Bone Marrow Edema and Synovitis in Patients with Spondyloarthritis Based on Magnetic Resonance Images. *Front. Physiol.* **2023**, *14*, 1132214. [\[CrossRef\]](#)
13. Adams, L.C.; Bressemer, K.K.; Ziegeler, K.; Vahldiek, J.L.; Poddubnyy, D. Artificial Intelligence to Analyze Magnetic Resonance Imaging in Rheumatology. *Jt. Bone Spine* **2024**, *91*, 105651. [\[CrossRef\]](#)
14. Jin, L.; Ma, Z.; Li, H.; Gao, F.; Gao, P.; Yang, N.; Li, D.; Li, M.; Geng, D. Interobserver Agreement in Automatic Segmentation Annotation of Prostate Magnetic Resonance Imaging. *Bioengineering* **2023**, *10*, 1340. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Moon, J.; Jadhav, P.; Choi, S. Deep Learning Analysis for Rheumatologic Imaging: Current Trends, Future Directions, and the Role of Human. *J. Rheum. Dis.* **2025**, *32*, 73–88. [\[CrossRef\]](#)
16. Omar, M.; Watad, A.; McGonagle, D.; Soffer, S.; Glicksberg, B.S.; Nadkarni, G.N.; Klang, E. The Role of Deep Learning in Diagnostic Imaging of Spondyloarthropathies: A Systematic Review. *Eur. Radiol.* **2024**, *35*, 3661–3672. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Momtazmanesh, S.; Nowroozi, A.; Rezaei, N. Artificial Intelligence in Rheumatoid Arthritis: Current Status and Future Perspectives: A State-of-the-Art Review. *Rheumatol. Ther.* **2022**, *9*, 1249–1304. [\[CrossRef\]](#)
18. Khan, R.F.; Lee, B.D.; Lee, M.S. Transformers in Medical Image Segmentation: A Narrative Review. *Quant. Imaging Med. Surg.* **2023**, *13*, 8747–8767. [\[CrossRef\]](#)
19. Schlereth, M.; Mutlu, M.Y.; Utz, J.; Bayat, S.; Heimann, T.; Qiu, J.; Ehring, C.; Liu, C.; Uder, M.; Kleyer, A.; et al. Deep Learning-Based Classification of Erosion, Synovitis and Osteitis in Hand MRI of Patients with Inflammatory Arthritis. *RMD Open* **2024**, *10*, e004273. [\[CrossRef\]](#)



20. Aizenberg, E. Computer-Aided Techniques for Assessment of MRI-Detected Inflammation for Early Identification of Inflammatory Arthritis. PhD Thesis, Leiden University, Leiden, The Netherlands, 2019.
21. Aizenberg, E.; Roex, E.A.H.; Nieuwenhuis, W.P.; Mangnus, L.; van der Helm-van Mil, A.H.M.; Reijnders, M.; Bloem, J.L.; Lelieveldt, B.P.F.; Stoel, B.C. Automatic Quantification of Bone Marrow Edema on MRI of the Wrist in Patients with Early Arthritis: A Feasibility Study. *Magn. Reson. Med.* **2018**, *79*, 1127–1134. [[CrossRef](#)]
22. Aizenberg, E.; Shamonin, D.P.; Reijnders, M.; van der Helm-van Mil, A.H.M.; Stoel, B.C. Automatic Quantification of Tenosynovitis on MRI of the Wrist in Patients with Early Arthritis: A Feasibility Study. *Eur. Radiol.* **2019**, *29*, 4477–4484. [[CrossRef](#)] [[PubMed](#)]
23. Kleinrensink, N.J.; Perton, F.T.; Pouw, J.N.; Vincken, N.L.A.; Hartgring, S.A.Y.; Jansen, M.P.; Arbabi, S.; Foppen, W.; De Jong, P.A.; Tekstra, J.; et al. TOFA-PREDICT Study Protocol: A Stratification Trial to Determine Key Immunological Factors Predicting Tofacitinib Efficacy and Drug-Free Remission in Psoriatic Arthritis (PsA). *BMJ Open* **2022**, *12*, e064338. [[CrossRef](#)] [[PubMed](#)]
24. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. Automated Design of Deep Learning Methods for Biomedical Image Segmentation. *Nat. Methods* **2019**, *18*, 203–211. [[CrossRef](#)] [[PubMed](#)]
25. Isensee, F.; Ulrich, C.; Wald, T.; Maier-Hein, K.H. Extending NnU-Net Is All You Need. *Inform. aktuell* **2023**, *2022*, 12–17. [[CrossRef](#)]
26. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
27. de Boer, M.; Kos, T.M.; Fick, T.; Van Doormaal, J.A.M.; Colombo, E.; Kuijff, H.J.; Regli, L.P.; Bartels, L.W.; Van Doormaal, T.P.C.; De Boer, M.; et al. NnU-Net versus Mesh Growing Algorithm as a Tool for the Robust and Timely Segmentation of Neurosurgical 3D Images in Contrast Enhanced T1 MRI Scans. *Acta Neurochir.* **2024**, *166*, 92. [[CrossRef](#)]
28. Pérez-García, F.; Sparks, R.; Ourselin, S. TorchIO: A Python Library for Efficient Loading, Preprocessing, Augmentation and Patch-Based Sampling of Medical Images in Deep Learning. *Comput. Methods Programs Biomed.* **2021**, *208*, 106236. [[CrossRef](#)]
29. Isensee, F.; Petersen, J.; Klein, A.; Zimmerer, D.; Jaeger, P.F.; Kohl, S.; Wasserthal, J.; Koehler, G.; Norajitra, T.; Wirkert, S.; et al. nnu-net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation. *arXiv* **2018**, arXiv:1809.10486.
30. Huang, L.; Miron, A.; Hone, K.; Li, Y. Segmenting Medical Images: From UNet to Res-UNet and NnUNet. In Proceedings of the 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), Guadalajara, Mexico, 26–28 June 2024; pp. 483–489. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.