Purpose: Automatic measurement of wrist cartilage volume in MR images.

Methods: We assessed the performance of four manually optimized variants of the U-Net architecture, nnU-Net and Mask R-CNN frameworks for the segmentation of wrist cartilage. The results were compared to those from a patch-based convolutional neural network (CNN) we previously designed. The segmentation quality was assessed on the basis of a comparative analysis with manual segmentation. The best networks were compared using a cross-validation approach on a dataset of 333 3D VIBE images of mostly healthy volunteers. Influence of some image parameters on the segmentation reproducibility was assessed.

Results: The U-Net-based networks outperformed the patch-based CNN in terms of segmentation homogeneity and quality, while Mask R-CNN did not show an acceptable performance. The median 3D DSC value computed with the U-Net_AL (0.817) was significantly larger than DSC values computed with the other networks. In addition, the U-Net_AL provided the lowest mean volume error (17%) and the highest Pearson correlation coefficient (0.765) with respect to the ground truth values. Of interest, the reproducibility computed using U-Net_AL was larger than the reproducibility of the manual segmentation. Moreover, the results indicate that the MRI-based wrist cartilage volume is strongly affected by the image resolution.

Conclusions: U-Net CNN with attention layers provided the best wrist cartilage segmentation performance. In order to be used in clinical conditions, the trained network can be fine-tuned on a dataset representing a group of specific patients. The error of cartilage volume measurement should be assessed independently using a non-MRI method.

KEYWORDS
arthritis, cartilage, deep learning, MRI, segmentation, wrist

Nikita Vladimirov and Ekaterina Brui contributed equally to this work.
1 | INTRODUCTION

Multiple morphological metrics computed from MR images have been reported with the aim of assessing the severity of inflammatory and degenerative diseases in articular joints. Quantitative MR-morphometry has been used for linear, surface, and volumetric measurements in bones,1–3 ligaments,4 meniscus,5 and articular cartilage.6–8 An accurate morphometric approach is expected to allow the detection of dynamic morphological changes related to the disease evolution or in response to a treatment so that the corresponding features could be used as imaging biomarkers.5,7,9 One has to keep in mind that the prerequisite for such a morphometric approach is related to an accurate delineation of the corresponding structures. This task, commonly performed manually, is tedious, time-consuming, and suffers from between- and within-operator variability.5,10

Over the past decades, MR morphometry has been used for large joints such as knee, hip, and shoulders11 likely because MR images of the corresponding joints had appropriate spatial resolution, contrast-to-noise ratio (CNR) and SNR. The accurate delineation of more complex and/or thinner joints from hand, wrist or feet requires images with a resolution that was not technically achievable until recently. Supportive of that, wrist cartilage assessment using MRI was not part of the Rheumatoid Arthritis MRI score until recently.12 Technical developments such as those in the field of radiofrequency coils have provided the opportunity to obtain images with a higher quality that could be used for a proper assessment of wrist cartilage.13,14 The initial segmentation approaches of wrist cartilage were manual with the commonly acknowledged caveats related to duration and reproducibility.5,15

More recently, automatic and semi-automatic segmentation methods have been reported for large structures such as bone,3,16,17 and very few of them have been reported for wrist cartilage likely as a result of the corresponding complex anatomy. An accurate segmentation method of wrist cartilage would be of high interest especially in rheumatoid arthritis (RA), in which the pathological process has been reported to start in small joints of hand18 and to affect cartilage rather than bone.19

The most promising methods for fully automatic segmentation of complex biomedical structures are based on convolutional neural networks (CNNs).20–22 Several network architectures have shown excellent results for cartilage segmentation from MR images of knee23 with Dice similarity coefficients (DSCs) up to 0.92,24,25 while a performance for wrist cartilage segmentation has been poorly reported. In a seminal study, Brui et al. reported a CNN patch-based approach dedicated to the automatic segmentation of wrist cartilage in 3D MR images.26 The planar CNN architecture provided an optimized segmentation of centrally-located slices with DSC values up to 0.81 whereas the corresponding segmentation for more lateral slices was poor thereby resulting in a low 3D segmentation accuracy (3D DSC = 0.69 ± 0.06). Overall, cartilage volume measurements over the whole 3D space was affected by large errors and other alternatives would be of high interest.

Fully convolution neural networks (FCNNs) have been largely used in the field of biomedical imaging. U-Net, the most popular architecture,24,26–30 is an autoencoder type network with skip-connections that transfer spatial contextual information from encoder to decoder, so that the contextual information of a whole image is merged with the final features thereby optimizing the spatial relationships recovery at the decoder level.30 U-Net has been already used for wrist cartilage segmentation and the corresponding results were poor (DSC = 0.64)10 likely because of skipping the step of hyperparameters optimization, which is a crucial step in CNN-based segmentation tasks. Optimization of parameters for CNN-based segmentation a critical issue for the performance of the corresponding CNN. Grid-search process has been commonly used for U-Net architectures whereas,24,26–30 in specific frameworks, such as nnU-Net,31 preprocessing, network architecture configuration, training, and post-processing procedures are proposed to be done automatically. Such an automatic procedure has shown promising performance in biomedical applications.32,33

Of interest, MR images of joints often contain structures such as skin and vessels, which display a contrast similar to cartilage so that they could be wrongly identified as cartilage.10 Attention layers29,34,35 added to a CNN architecture have been developed so that the CNN can learn from relevant regions only. Such an approach has provided interesting results for the delineation of pancreas34 and for the segmentation of knee menisci.29 Such an approach has never been reported for wrist cartilage segmentation. Another alternative relies on a preliminary detection of the object of interest. Among the proposed tools, Mask R-CNN,36 an instance segmentation framework, can be considered as the state-of-the-art for medical image segmentation tasks and has shown promising results in musculoskeletal37 and breast cancer segmentation.38

In the present study, we intended to assess whether wrist cartilage segmentation can be improved by using state-of-the-art CNNs, such as U-Net (including U-Net with attention layers), nnU-Net, and Mask R-CNN. The performance of these CNNs was compared to the patch-based CNN we initially reported.10 The networks, which provided the best performance, were also trained and tested on an extended dataset. Their performance was thoroughly assessed on the basis of both conventional
segmentation metrics such as DSC and cartilage volume measurements. The source code is available at GitHub.39

2 | METHODS

2.1 | Datasets and subjects

The study was approved by the local ethics committee of Federal Almazov North-West Medical Research Center. Two datasets referred as “small” (SDS) and “large” (LDS) were used. The SDS was used for an initial comparative analysis regarding segmentation homogeneity within the wrist volume among the tested networks and a previously proposed PB-CNN.10 The SDS consisted of 560 2D slices selected (every other slice) from 20 3D MR images of wrist recorded in 11 subjects (8 healthy and 3 with confirmed osteoarthritis [OA] or RA).10 This dataset contained MR images obtained from a single 1.5T MRI scanner and a single imaging sequence (3D VIBE). It was halved in two subsets, one for training (50%) and one for testing (50%), in order to perform a hold-out training. The validation dataset used during the training session was set to 10% of the training dataset.

Given that the homogeneity of SDS could be considered as a limitation, we extended the SDS by adding MR images acquired with another scanner (3T) and also images obtained with the same scanner but with variable acquisition parameters, including voxel size (see the details in Supporting Information). This enriched dataset was referred as the LDS. Images from the LDS (1297 2D slices) were those from the SDS plus 10 3D MR images recorded in five subjects (four healthy volunteers and one patient with a confirmed RA). Some subjects were scanned twice and some had a scan of both hands. A cross-validation training approach was used in order to estimate the achievable performance of the different CNNs on heterogeneous data. Using the GroupKFold from sklearn library,40 the LDS was divided into five subsets (with 20% of the total amount of 3D images in each) and used for a five-fold cross-validation analysis.

Both datasets were manually labeled by a more than 10-y experienced radiologist as previously indicated10 (V.F.). Training and testing subsets did systematically contain data from different subjects. For the training session, images size was standardized (central cropped) to 256×256. For the training subsets, an augmentation step was performed using albumentations library41 and included vertical and horizontal flip, arbitrary angle rotation, elastic transformation and grid distortion. Each image was nine-times augmented, resulting in addition of nine augmented copies for each image into the initial datasets dataset. The performed augmentations had the following value ranges: vertical flip probability = 0.7, horizontal flip probability = 0.7, rotate limit = 180° and probability = 0.6, elastic transform alpha = 10, sigma = 30, alpha_affine = 25, probability = 0.4, grid distortion num_steps = 15, distort_limit = 0.2 and probability = 0.4. Images were normalized with respect to the mean intensity and standard deviation. Each dataset was represented by a set of 2D matrices: \( D = (X_i, Y_i)_{i=1}^N \), where \( X_i \) is a 2D MR image of wrist and \( Y_i \) – the corresponding 2D cartilage binary mask.

2.2 | Convolutional neural networks

All the networks were trained on a high-performance cluster with the following characteristics: CPU: 2× Xeon Silver 4214, RAM: 128 Gb, GPU: 4× T V100. CNN development was performed using Python programming language and TensorFlow and Keras open source libraries. Training was performed on all available GPUs.

2.2.1 | U-Net-based networks

We initially used a classical U-Net architecture. Layers of batch normalization, noise and spatial dropout were added in order to improve convergence time and decrease generalization error. The noise was added as a TensorFlow layer directly to the CNN input. In addition to the training hyperparameters and as previously described,22,42 number of layers (i.e., the depth of the CNN), order and presence of noise, dropout and batch normalization layers have been considered as adjustable architecture hyperparameters. Two variants of the U-Nets were used: one with an original depth, that is, number of max-poolings equal to 4,30 and one with a reduced depth (Truncated U-Net) that is, with three max-poolings. The order and the presence of dropout and batch normalization layers in different parts of the networks were adjusted so as to maximize the averaged 3D DSC during the testing phase. Overall, we assessed 12 configurations for both U-Net and Truncated U-Net. An Adam optimizer was used for training with a fixed batch size (32) and cross-entropy was used as the loss function. Several hyperparameters were adjusted throughout the training process via a grid search that is, learning rate (from \( 6 \times 10^{-4} \) to \( 3 \times 10^{-3} \)), utilization of learning rate decay (from \( 50 \times 10^{-4} \) to \( 5 \times 10^{-4} \)), noise level (from 0 to 0.5), and dropout probability (from 0 to 0.5).

With the aim of avoiding false positive results from regions out of the wrist joint area (an issue reported in a previous work10), attention layers were added in the U-net and Truncated U-Net. A localization layer was integrated to the skip connections as previously proposed.34
Briefly, feature maps from encoder and decoder propagate through $1 \times 1$ size aligning convolutions and are added so that the resulting feature map values are modulated according to the feature map values of both sources. Then, feature maps are activated by a ReLU layer and normalized by a sigmoid function. The attention coefficients are then applied after a bilinear upsampling. Thus, the attention layers are intended to compute attention coefficients that are multiplied with feature maps in an element-wise manner. In that way, less informative regions can be ignored. The structure of an attention layer depicted in Figure S1 illustrates the corresponding operation in details.

We also investigated the performance of a recently proposed framework that is, nnU-Net.\textsuperscript{31} nnU-Net (no-new U-Net) is a framework with predefined optimization of fixed parameters, rule-based parameters and empirical parameters. Fixed parameters are related to learning rate schedule, loss function, architecture template, optimizer, data augmentation, training and inference procedures. Among these options, we have excluded data augmentation procedure, in order to make the comparison with other networks more consistent, as the data had been initially augmented. In addition, based on prior knowledge, we have preselected cross-entropy as a loss function. The rule-based parameters are related to dataset fingerprints and pipeline characteristics, such as network topology, patch and batch size, and so forth. Their optimization is based on heuristic expert knowledge and is aimed to achieve optimal GPU memory utilization and a sufficiently large receptive field. The empirical parameters are related to model selection and post-processing, and are derived from analyzing the performance of the trained networks.

### 2.2.2 Mask R-CNN

In addition to the above mentioned architectures, we investigated the performance of Mask R-CNN.\textsuperscript{36} The initial MR images and manually segmented masks were resized (upscaled) to $512 \times 512$. Mask R-CNN provides an instance segmentation. In that respect, the data labeling step had to be enriched with the instance and bounding box labeling. We tested two approaches for instance definition during the labeling process. As a first approach, we considered the whole set of cartilage pixels from a single slice as part of a given instance surrounded by a bounding box (single-instance labeling). As a second approach, considering the complex anatomy of wrist cartilage, each image was considered as a multi-instance object (multi-instance labeling). In other words, masks were additionally processed using a home-built Python script so as to split them in multiple instances. The splitting algorithm is provided in Figure S2. We added a 10 pixel overlap region between bounding boxes so as to provide a smooth junction between neighboring instances.

Mask R-CNN contains an enormous number of adjustable parameters. For most of the parameters we used default values provided in.\textsuperscript{43} Two variants of a backbone network were tested: ResNet-50 and ResNet-101. We also utilized the provided pretrained weights (MS COCO) for all networks included into Mask R-CNN, and fine-tuned the last layers only. The models were trained for 100 epochs with a learning rate of 0.001. Mask R-CNN utilizes a complex loss function, which combines class label prediction, bounding box refinement, and segmentation mask prediction loss functions with adjustable weights. We hypothesized that with different number of instances, the optimal weights for these loss functions should be different, and thus tested different weights for segmentation mask prediction loss function (in a range from 1 to 5, increment 1), while the weights for combined class label prediction and bounding box refinement loss functions were fixed at the value of 1. These weights were adjusted so as to maximize the averaged 3D DSC during the testing phase.

### 2.3 Data analysis

#### 2.3.1 Metrics

Several metrics were used in order to assess the network’s performance. As a main metric we used DSC:\textsuperscript{25}

$$DSC = \frac{2 \cdot |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|}$$

where $\hat{Y}$ is a binary cartilage mask predicted by the CNN, and $Y$ - a ground-truth (GT) mask obtained by a human observer. DSC was calculated for each individual slice (2D DSC) and also for each of 3D images (3D DSC) in the testing datasets. In addition, an overall precision metric was computed as follows:\textsuperscript{44}:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where $TP$ is an overall number of true positive pixels in the predicted masks within the whole testing set, $FP$ - a corresponding number of false positive, pixels.

As previously described,\textsuperscript{10} a region-based analysis was used to assess the CNNs’ performance across the wrist joint in terms of mean 2D DSC over particular wrist zones. The 2D DSC values distribution was analyzed with respect
to cartilage representation. Masks of the slices were gathered in four bins according to the relative amount of cartilage in the slice as compared to the central slice, that is, the slice with the highest amount of cartilage pixels (1, 0%; 2, 0%–33%; 3, 34%–66%; 4, 67%–100%), and the corresponding mean DSC values were computed in these zones. In addition to DSC, cartilage volume was computed given that it is a metric of interest in joints disorders. 6 The error of cartilage volume ($\Delta V$) measurement was computed as follows:

$$\Delta V = \frac{|V_{GT} - V_{pred}|}{V_{GT}} \cdot 100\%$$

where $V_{GT}$ is a GT cartilage volume computed from a manual segmentation and $V_{pred}$ is the corresponding volume computed from the automatic segmentation from a given CNN.

### 3.2 | CNN performance among the zones

As illustrated in Figure 1 and as previously described, the results obtained with the PB-CNN differed significantly according to the zones. 2D DSC in zone 1 was 0.210 whereas it was significantly larger for zones 2–4 (DSC ranged from 0.600 to 0.730). Both models of Mask R-CNN provided an opposite trend: the DSC in zone 1 was impressively high when the CNN was trained using single instance approach (0.940) and lower in case of
multi-instance approach (0.891). At the same time, their performance in zones 2–4 was poor (from 0.523 to 0.673). On the contrary, based on the DSC values, performance of all U-Net-based CNNs was less zones-dependent. The most significant improvement with respect to the PB-CNN was observed for zone 1 using U-Net (from 0.210 to 0.919). The large DSC standard deviations calculated for zone 1 can be explained by the fact that, as this zone did not contain the cartilage in the GT masks, the similarity coefficient with the predicted masks could be equal either to 0 (if the network segmented false positive pixels) or 1 (if the network correctly did not segment anything). Interestingly, both models of Mask R-CNN provided relatively small standard deviations (0.08 – for single instance training and 0.1 – for multi-instance) in zone 1, thereby indicating an excellent ability for detection of regions containing wrist cartilage.

Results of the overall 3D CNN performance are summarized in Table 1. As compared to the PB-CNN, all the U-Net-based CNNs provided a higher 3D DSC value, with the U-Net_AL providing the highest value (0.811 ± 0.037). However, ANOVA test did not indicate any significant influence of the U-Net-based CNN variant on the mean

---

**TABLE 1** 3D DSC values and time performance on the Small Dataset (SDS).

|              | PB-CNN 10 | U-Net | U-Net_AL | Tr-U-Net | Tr-U-Net_AL | nnU-Net | Mask R-CNN (SI) | Mask R-CNN (MI) |
|--------------|-----------|-------|----------|----------|-------------|--------|----------------|----------------|
| 3D DSC       | 0.690 ± 0.060 | 0.798 ± 0.076 | 0.811 ± 0.037 | 0.802 ± 0.035 | 0.805 ± 0.035 | 0.806 ± 0.042 | 0.50 ± 0.045 | 0.655 ± 0.036 |
| Trainable parameters | 73 × 10^3 | 31 × 10^6 | 32 × 10^6 | 7.7 × 10^6 | 8.0 × 10^6 | 30 × 10^6 | 64 × 10^6 | 64 × 10^6 |
| Processing time (s)a | >60 | 3.32 | 3.47 | 2.31 | 2.44 | 2.20 | 26 | 26 |
| Learning time  | 74.4 h b | 67.2 min | 68.6 min | 54.0 min | 53.0 min | 4.5 h | 11.3 h | 11.9 h |

Abbreviations: MI, multi-instance; SI, single instance.

aData were processed using a conventional computer with Nvidia GTX 1050 2 Gb, batch size = 2.

bThe training was performed on a server with four processors (Intel Xeon E5-4617 2.90 GHz) and 512 Gb RAM.
Both variants of Mask R-CNN had a poor 3D performance even in comparison to PB-CNN. Duration of the CNNs training with the SDS and the processing time for a single dataset (a 3D image with 88 slices) are also summarized in Table 1. All the U-Net-based networks outperformed the PB-CNN with a ~20 times reduction for the processing time. The faster network was nnU-Net in terms of processing time (2.20 s), and Tr_U-Net_AL in terms of training time (53 min).

### 3.3 Cross-validation on the LDS

For the LDS, only the U-Net-based models were used given their similarly high performance. 2D and 3D metrics were computed for all these CNNs with an extended dataset intended with the aim of assessing the CNNs performance on a more heterogeneous data in terms of image characteristics. Using the cross-validation approach, we made sure to estimate the best achievable characteristics for the CNNs.

The largest median 2D DSC was acquired for U-net_AL network (see the boxplots in Figure 2 and the first line in Table 2).

![Figure 2](image-url)  
**Figure 2** Distributions of 2D DSC (merged scatter plots and boxplots), for the CNNs trained and tested on a big dataset (outliers, i.e., $>Q3 + 1.5 \times \text{IQR}$ or $<Q1-1.5 \times \text{IQR}$, where IQR is an interquartile range, are excluded from the analysis).

Volumetric DSCs for all the studied CNNs are presented in Table 2 together with the precision values. The one-way ANOVA illustrated that the 3D DSC values were influenced by the type of network ($f(2) = 2.03, p = 0.09$). Post hoc tests indicated a significant difference between U-Net and U-Net_AL CNNs ($p = 0.007$) and between nnU-net and U-Net_AL CNNs ($p = 0.09$). U-Net_AL showed the largest 2D and 3D DSC median values that is, 0.822 and 0.817. The largest precision value was quantified for nnU-Net (0.810). Networks with attention layers showed a higher precision as compared to the corresponding networks without attention layers.

Figure 3 provides the results of the CNNs testing on the slices selected from different zones of a 3D image and illustrates the architecture choice effect. One can note an increase for 2D DSC index and a reduction of the number of false positive pixels in case of adding the attention layers to the U-Net network.

### 3.4 Cartilage volume measurements

A 3D visualization of cartilage is provided in Figure 4. The average volume measurements computed from the manual and the automatic segmentations are summarized in Table 3. No significant CNN-effect was observed for the 3D wrist cartilage volume ($f(2) = 0.52, p = 0.72$). As indicated in Table 3, the mean relative volume errors ranged from 17.21% (U-Net_AL) to 21.40% (nnU-Net). Correlations between volumes computed manually and automatically are illustrated in Figure 5. The highest Pearson correlation coefficient was observed for U-Net_AL ($r = 0.765$) (Figure 5A). Bland–Altman plots displayed in Figure 5(B) indicated that the smallest volume difference range ($\pm 1.96\sigma$) (green and blue solid lines) was achieved for U-Net_AL. Overall, the volume difference was independent of the volume (Figure 5B) in all cases. For all the networks, the systematic bias ranged from 1.6 (Tr_U-Net_AL) to 164.7 (nnU-Net) mm$^3$. As the individual biases varied in wide ranges (in general, larger

| Metric   | U-Net | U-Net_AL | TrU-Net | TrU-Net_AL | nnU-Net |
|----------|-------|----------|---------|------------|---------|
| 2D DSC   | 0.808 | 0.822    | 0.817   | 0.815      | 0.81    |
|          | [0.739, 0.868] | [0.755, 0.876] | [0.744, 0.864] | [0.744, 0.865] | [0.702, 0.885] |
| 3D DSC   | 0.810 | 0.817    | 0.799   | 0.811      | 0.814   |
|          | [0.780, 0.822] | [0.785, 0.838] | [0.775, 0.828] | [0.784, 0.823] | [0.748, 0.849] |
| Precision| 0.773 | 0.780    | 0.769   | 0.781      | 0.810   |

Note: For the DSC, values are reported as median and range (25th percentile, 75th percentile). Precision is estimated within the whole testing set.
Example of CNNs performance in different zones of wrist. Red, false positive; blue, false negative; and green, true positive pixels.
than \([−1000 \, 1000] \, \text{mm}^3\), the systematical bias was very sensitive to a small size of the sample. Thus, we considered the \(±1.96σ\) as a more reliable metric of network performance quality.

In the LDS, seven subjects were scanned repeatedly in a single session using two different coils, one of them providing higher and lower SNR.\(^{49}\) The absolute difference for the cartilage volume was computed and scaled (%) to the largest value in the pair using the values from repeated scans (Table 4). The difference was lower for U-Net_AL in most of the cases with values ranging from 0.2% to 16.6%. Table 4 also contains the amounts of false negative and false positives pixels in repeated measurements. No systematic trend from statistical point of view was observed.

There were three cases in the LDS when the same hand of the same subject had been scanned twice with different resolution. The cartilage volume difference measured from the images with different resolution is summarized in Table 5. In all three cases, the raise of the voxel size led to an increase of the measured volume.
TABLE 4  Results for repeated measurements.

| Subject  | 1     | 2     | 3     | 4     | 5     | 6     | 7     |
|----------|-------|-------|-------|-------|-------|-------|-------|
| Manual   | ΔVrep,% | 13.9  | 11.8  | 8.6   | 1.9   | 28.2  | 8.4   | 16.5  |
| FP       | 14 594| 10 965| 20 518| 19 403| 9394  | 17 632| 36 042|
| FN       | 17 657| 10 977| 21 874| 24 565| 12 584| 18 740| 37 531|
| U-Net + AL| ΔVrep,% | 1.6   | 16.6  | 3.5   | 1.9   | 16.3  | 8.9   | 0.2   |
| FP       | 18 779| 13 449| 24 153| 16 710| 12 587| 20 832| 30 201|
| FN       | 19 392| 12 510| 22 830| 21 181| 9946  | 21 981| 26 152|

Note: Cartilage volume difference was measured from MR images segmented manually or with the U-Net-AL from the pairs of repeated scans. Repeated images were obtained with a home built wrist coil (higher SNR) and with a commercial extremity coil (lower SNR).

Abbreviations: ΔVrep, cartilage volume difference; FP, number of false positive; FN, false negative pixels from repeated measurements.

TABLE 5  Cartilage volume difference measured from MR images with different voxel size, %.

| Subject  | 1     | 2     | 3     |
|----------|-------|-------|-------|
| Voxel change, mm³ | 0.146×0.146×0.4 | 0.318×0.316×0.5 | 0.391×0.391×0.5 |
| Manual   | 25.87 | 5.98  | 18.28 |
| U-Net + AL| 0.09 | 19.82 | 12.36 |

Note: Cartilage volume was measured from MR images segmented manually or with the U-Net-AL. The same hand of the same subject was scanned twice with different resolution.

4 | DISCUSSION AND CONCLUSION

In the present work, various convolutional neural networks have been used for wrist cartilage segmentation and volumetric measurements, and the corresponding results have been compared. We hypothesized that the image-based or instance-based approaches could be more efficient than the patch-based approach we initially tested, and for which only the centrally located slices were properly segmented while a large number of FP pixels was predicted in lateral slices and outside the joint area. Four variants of U-Net CNN with different depth and manually optimized hyperparameters have been tested. An attention mechanism was integrated to the skip connections of the U-Nets in order to increase the ability of the networks to ignore the area of the image outside the wrist joint. In addition nnU-Net – a U-Net with automatic adjustment of preprocessing, network architecture configuring, training and post-processing procedures was tested. Overall, for the same dataset, all the U-Net based CNNs outperformed the patch-based CNN in terms of segmentation homogeneity and quality all over the wrist volume, including background-only slices. These results were partially presented at ISMRM-2022. Using an extended dataset, the U-Net architecture with additional attention layers provided the best results, with a 0.817 3D DSC value and a mean relative error of cartilage volume of 17.21%. An instance-based approach – Mask-R-CNN – also drastically reduced false positive predictions outside the wrist cartilage location (zone 1). However, this approach was not beneficial for cartilage areas.

4.1 | U-Net-based CNNs

The results computed from the SDS indicate a clear benefit of image-based learning approach as compared to the previously reported patch-based strategy. Patch-based approach for images segmentation has been reported as of interest given that a more comprehensive data augmentation by multiple patch views and class balancing can be performed. According to the present comparative analysis, these features are not adapted for an optimal 3D segmentation of wrist cartilage. It has been indicated that a PB-CNN approach could be effective as long as patches could contain enough contextual information. Our results indicate that this information would not be homogenous and then not enough informative. We have already reported an issue related to generalization of a patch-based learning approach. More specifically, cartilage from the lateral slices was not properly
segmented likely because patches from these slices did contain a contextual information largely different than patches positioned in centrally located slices. One has to keep in mind that, in addition to that, patches with “lateral” cartilage were weakly represented in the training dataset due to the relatively smaller amount of cartilage in lateral slices. The tested U-Net-based CNNs were not only more efficient regarding the segmentation metrics but also regarding the training time. Accordingly, it has been reported that patch-based approaches were less computationally effective than fully convolution networks with an image-based learning approach as they classify each pixel in an image separately. It is important to mention that, although the training time for the U-Net optimized in nnU-Net framework was longer than for the common U-Nets, the overall time the authors spent for manual adaptation of the networks and for the grid search was much longer. Overall, our results are clearly supportive of the better performance of image-based CNNs as compared to patch-based approaches and extend this to 3D wrist cartilage segmentation.

Using an extended dataset that is, the LDS, the U-Net-based CNNs performance was compared on the basis of a cross-validation approach. This comparative analysis was based on both geometric and volumetric measurements. Even though the 3D DSC values were very close to each other, the statistical analysis indicated that the median 3D DSC computed with the U-Net_AL was significantly larger. In addition, the U-Net_AL provided the lowest mean volume error, smallest volume difference range (±1.96σ), and the highest Pearson correlation coefficient with respect to the GT. Even though the cartilage volumes predicted by all the U-Net based CNNs were not statistically different from the other networks, these results illustrate that attention layers added to the U-Net allowed to achieve the best wrist cartilage segmentation. On the contrary, the truncation of the U-Net did not modify the segmentation performance. Interestingly, manual adjustment of the U-Net parameters was more effective in this particular task than their automatic adaptation in nnU-Net framework. Configuration adaptations provided by nnU-Net are based on execution of a set of rules pre-defined by a prior knowledge about the best experiences. Thus, being a truly remarkable tool, nnUnet framework does not execute a comprehensive optimization of the architecture, that can be required in some particular cases.

Attention layers can minimize signal from non-contributive regions and maximize it in the regions of interest. Overall, this attention mechanism is expected to reduce false positive predictions. As expected, addition of attention layers to the U-Nets was linked to an increased 3D DSC (from 0.810 to 0.817) and a reduced volume error (from 19.72% to 17.21%) while the segmentation accuracy was also raised. In the present study, the inclusion of attention mechanism allowed to improve the 3D performance although this improvement was not very critical.

It is important to mention that a large variety of attention mechanisms have been reported so far but the corresponding added value was controversial. Using U-Net with Multi Global Attention, Dayananda et al. reported a 3% DSC improvements as compared to the U-Net architecture in brain segmentation. Tong et al. (cardiac segmentation) proposed a two-stage framework with an initial detection process followed by a segmentation combined to a complex interleaved attention module that combines both spatial and channel attention. Although this architecture was expected to keep a significant amount of spatial and semantic information, the DSC improvement as compared to the U-Net architecture was 1% only. Zhang et al. (brain tumor segmentation) incorporated U-Net skip connections with attention layers and ResBlocks and reported a modest DSC improvement that is, 1%. Küstner et al. (adipose tissue segmentation) proposed blending U-Net, ResNet, and DenseNet, with the latter enabling the use of merge-and-run mappings for multiresolution attention. A very significant (47%) DSC improvement was reported as compared to U-Net. However, one must keep in mind the poor performance of the initial U-Net-based segmentation (with average DSC = 0.64). Overall, our results indicating a 1.7% improvement in 2D DSC using basic attention mechanisms are very similar to those reported in the literature. Further work is warranted in order to reach significant beneficial effects of attention mechanisms for image segmentation tasks.

4.2 Mask-RCNN

Mask R-CNN is a state-of-the-art framework for object detection and instance segmentation, which has shown impressive performance in instance segmentation on different domain, including medical image analysis. Mask R-CNN consist of a backbone network for feature extraction, a regional proposal network, a branch for prediction of bounding box and class of the proposed object, and a parallel branch for prediction of object mask. Due to its ability and high efficiency in instance detection and further segmentation, we expected some beneficial effects for wrist cartilage segmentation and more particularly a decrease of the false positive rate. The excellent results obtained for zone 1 definitely confirmed this assumption. However, in terms of 3D DSC, training Mask R-CNN on the SDS, while considering the whole cartilage as one instance, did not provide any improvement. We then divided a given image in multiple instances considering the complexity of
the cartilage anatomy. The corresponding 2D DSC (0.655) was poor but comparable to previous results obtained for knee femoral cartilage (2D DSC = 0.71) and effusion segmentation (2D DSC = 0.71). Although we did not achieve an acceptable performance using Mask R-CNN for the wrist cartilage segmentation, further work toward a better multi-instances data labeling might provide better results.

4.3 Sources of errors

In degenerative and inflammatory diseases of joints, cartilage loss is considered as a quantitative metric of cartilage damage. Assessment of wrist cartilage volume has not been reported to date, likely as a result of the complex anatomy, the cartilage size, and the difficulties related to segmentation. Such a quantitative assessment would be of interest as long as the corresponding accuracy and volume error are adequate to changes reported so far in pathological situations. To the best of our knowledge, wrist cartilage volume changes in OA or RA have not been reported so far. According to studies conducted over a 10-y period in knee of osteoarthritic patients, the cartilage loss in the medial and lateral compartments would be 19.1% and 13.8% respectively. On that basis, the volume error we obtained (17.2%) with U-Net_AL should be adequate for the detection of cartilage loss in arthritis.

One has to keep in mind that our geometrical and volumetric indices were computed with respect to the manual segmentation, which is inherently affected by errors. We previously reported a 0.90 DSC value for repeated manual segmentations of wrist cross-sectional slices by the same observer. In the present study, we had the opportunity to compute cartilage volume errors from repeated MRI scanning. Interestingly, the corresponding results disclosed a lower reproducibility of manual cartilage volume measurements as compared to the U-Net_AL-based quantification. This may indicate that the network that has been trained on a heterogeneous data, both in terms of image quality and manual labeling quality, provides a more reproducible result than the human observer. Of interest, this illustrates an optimal generalization ability of the CNN. In addition, U-Net_AL performed consistent in terms of FP and FN. Another important issue in cartilage volume quantification is related to image resolution. On the basis of microcomputed tomography, it has been reported that cartilage thickness in a healthy wrist joint varied according to location from 0.6 to 1 mm. In our study, the voxel size was within this range, that is, between 0.146 × 0.146 × 0.4 mm³ and 0.508 × 0.508 × 0.5 mm. Although the sample size was small for averaging and generalizing the results, the results obviously showed that the MRI-based wrist cartilage volume is affected by the image resolution. These findings indicated that, even though the manual labels are commonly considered as the GT, they should be questioned as the “gold standard.” For this reason, a more precise validation of wrist cartilage volume in future should rely on independent non-MRI-based measurements, for example, direct measurement of cartilage volume in cadaver joints.

4.4 Limitations

We have to acknowledge that the datasets used in the present study were mainly composed of healthy volunteers with a few patients only. On that basis, for a potential clinical application, one could have to fine-tune the network reported in the present study. As a follow up of the present study, it could be of interest to perform multiclass segmentation for cartilage of each carpal bone similarly to what has been done for knee cartilage. Mask R-CNN, atlas-based, graph-based, multiclass CNN-based tools could be useful for this task. We also acknowledge that the distribution of parameters used for data acquisition, was not homogeneous (2/3 of the LDS had a voxel size of 0.37x0.37x0.5 mm³ and the other 1/3 contained images with various voxel size). However, we have carefully controlled that the images of different voxel size were similarly represented in both training and test samples during cross-validation.

U-Net convolution neural networks provided a significantly higher segmentation homogeneity within a 3D wrist VIBE image than a previously proposed PB-CNN. U-Net with additional attention layers provided the best segmentation quality. In order to be used in clinical conditions, the network can be fine-tuned on a dataset representing a group of specific patients. The error of cartilage volume measurement should be assessed independently using a non-MRI method in order to estimate the accuracy of the method. These additional studies will strengthen the obtained results and potentially further improve the segmentation quality.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education and Science of the Russian Federation (075-15-2021-592). Part of the work related to Mask R-CNN was supported by a grant for scientific school HIII-2359.2022.4.

ORCID

Nikita Vladimirov © https://orcid.org/0000-0003-1943-9139
REFERENCES

1. Eckstein F, Burstein D, Link TM. Quantitative MRI of cartilage and bone: degenerative changes in osteoarthritis. NMR Biomed. 2006;19:822-854. doi:10.1002/nbm.1063

2. Li X, Yu A, Virayavanich W, Noworolski SM, Link TM, Imbojj J. Quantitative characterization of bone marrow edema pattern in rheumatoid arthritis using 3 tesla MRI. J Magn Reson Imaging. 2012;35:211-217. doi:10.1002/jmri.22803

3. Aizenberg E, Roex EAH, Nieuwenhuis WP, et al. 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. doi:10.1002/mrm.26712

4. Buck RJ, Wyman BT, Hellio Le Graverand MP, Hudelmaier M, Wang J, de Herrera AGS. A review on segmentation of knee articular cartilage: from conventional methods towards deep learning. Artif Intell Med. 2020;106:101851. doi:10.1016/j.artmed.2020.101851

5. Haraldsen G, Aletaha D, Funovits J, Smolen JS. Physical disability in rheumatoid arthritis is associated with cartilage damage rather than bone destruction. Ann Rheum Dis. 2011;70:733-739. doi:10.1136/ard.2010.138693

6. Foster B, Joshi AA, Borgese M, Abdelhafef Y, Boutil RD, Chaudhari AJ. WRIST: a WRist Image Segmentation Toolkit for carpal bone delineation from MRI. Comput Med Imaging Graph. 2018;63:31-40. doi:10.1016/j.compmedimag.2017.12.003

7. Mohammed RH, Goyal A, Bansal P. Hand and Wrist Rheumatoid Arthritis. StatPearls; 2022. https://www.ncbi.nlm.nih.gov/books/NBK560890/

8. Zink JV, Souteyrand P, Guis S, et al. Standardized quantitative measurements of wrist cartilage in healthy humans using 3 T magnetic resonance imaging. World J Orthop. 2015;6:641-648.

9. Buck RJ, Wyman BT, Hellio Le Graverand MP, Hudelmaier M, Wirth W, Eckstein F. Osteoarthritis may not be a one-way road – comparison of spatial patterns of cartilage change between osteoarthritic and healthy knees. Osteoarthritis Cartilage. 2010;18:329-335. doi:10.1016/j.joca.2009.11.009

10. Eckstein F, Wirth W. Quantitative cartilage imaging in knee osteoarthritis. Rheumatology. 2021;60:1392-1399. doi:10.1093/rheumatology/keaa522

11. Neogi T, Bowes MA, Niu J, et al. Magnetic resonance imaging-based three-dimensional bone shape of the knee predicts onset of knee osteoarthritis: data from the osteoarthritis initiative. Arthritis Rheum. 2013;65:2048-2058. doi:10.1002/art.37987

12. Brui E, Efimtcev AY, Fokin VA, et al. Deep learning-based fully automatic segmentation of wrist cartilage in MR images. NMR Biomed. 2020;33:e4320. doi:10.1002/nbm.4320

13. Simonini FG, Cericin M, Cimaz R, et al. Evidence for immune activation against oxidized lipoproteins in inactive phases of juvenile chronic arthritis. Journal Rheumatol. 2001;28:198-203.

14. Herz B, Albrecht A, Englbrecht M, et al. Osteitis and synovitis, but not bone erosion, is associated with proteoglycan loss and microstructure damage in the cartilage of patients with rheumatoid arthritis. Ann Rheum Dis. 2014;73:1101-1106.

15. McQueen FM, McHaffie A, Clarke A, et al. MRI osteitis predicts cartilage damage at the wrist in RA: a three-year prospective 3 T MRI study examining cartilage damage. Arthritis Res Ther. 2014;16:R33. doi:10.1186/ar4462

16. Peterfy CG, van Dijke CF, Lu Y, et al. Quantification of the volume of articular cartilage in the metacarpophalangeal joints of the hand: accuracy and precision of three-dimensional MR imaging. Am J Roentgenol. 1995;165:371-375. doi:10.2214/ajr.165.2.7618560
carotid artery stenosis and carotid atherosclerotic plaque evaluation. *Front Physiol.* 2022;6:1057800. doi:10.3389/fphys.2022.1057800

33. Pettit RW, Marllatt BB, Corr SJ, Havelka J, Rana A. nnU-net deep learning method for segmenting parenchyma and determining liver volume from computed tomography images. *Ann Surg Open.* 2022;3:e155. doi:10.1097/ASO.0000000000000155

34. Oktay O, Schlemper J, le Folgoc L, et al. Attention U-net: learning where to look for the pancreas. *ArXiv.* 2018. doi:10.48550/arXiv.1804.03999

35. Chen L, Zhang H, Xiao J, Nie L, Shao J, Chua TS. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017:6298-6306. doi:10.1109/CVPR.2017.667

36. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell.* 2020;42:386-397. doi:10.1109/TPAMI.2018.2844175

37. Felfeliyan B, Hareendranathan A, Kuntze G, Jaremko JL, Ronsky JL. Improved-mask R-CNN: towards an accurate generic MSK MRI instance segmentation platform (data from the osteoarthritis initiative). *Comput Med Imaging Graph.* 2022;97:102056. doi:10.1016/j.compmedimag.2022.102056

38. Zhang Y, Chan S, Park VY, et al. Automatic detection and segmentation of breast cancer on MRI using mask R-CNN trained on non-fat-sat images and tested on fat-sat images. *Acad Radiol.* 2022;29:S135-S144. doi:10.1016/j.acra.2020.12.001

39. Vladimirov N. Github: wrist-segmentation. https://github.com/vnikale/wrist-segmentation.

40. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res.* 2011;12:2825-2830.

41. Buslaev A, Parinov A. Albuttomerations: fast and flexible image augmentations. *Information.* 2018;11(2):125. doi:10.3390/info11020125

42. Isensee F, Petersen J, Klein A, et al. nnU-Net: self-adapting framework for U-Net-based medical image segmentation. *Nat Methods.* 2021;18:203-211. doi:10.1038/s41592-020-01008-z

43. Matterport, Mask_RCNN, GitHub. https://github.com/matterport/Mask_RCNN 2017.

44. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020. doi:10.48550/arXiv.2010.16061

45. Stihle L, Wold S. Analysis of variance (ANOVA). *Chemometr Intell Lab Syst.* 1989;6:259-272. doi:10.1016/0169-7439(89)80095-4

46. ANOVA in R-Stats and R. https://statsandr.com/blog/anova-in-r/#post-hoc-test. Accessed May 10, 2022.

47. Bland JM, Altman GD. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1:307-310. doi:10.1016/S0140-6736(86)90837&hyphen;8

48. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J.* 2012;24:69.

49. Brui E, Mikhailovskaya A, Solomakha G, Efimtcev A, Andreychenko A, Shchelokova A. Volumetric wireless coil for wrist MRI at 1.5 T as a practical alternative to Tx/Rx extremity coil: a comparative study. *J Magn Reson.* 2022;339:107209. doi:10.1016/J.JMR.2022.107209

50. Vladimirov N, Brui E, Levchuk A, Efimtcev A, Bendahan D. Wrist cartilage segmentation using U-net convolutional neural networks enriched with attention layers. *Soc Mag Reson Med.* 2022;30:2969.

51. Zhang C, Cheng J. Image scoring: patch based CNN model for small or medium dataset. 2017 3rd IEEE international conference on computer and communications, ICCC 2017. 2018. doi:10.1109/COMPCOM.2017.8322898

52. Long J, Shollamer E, Darrell T. Fully convolutional networks for semantic segmentation. IEEE conference on computer vision and pattern recognition (CVPR). 2015.

53. Dayananda C, Choi JY, Lee B. Multi-scale squeeze U-SegNet with multi global attention for brain MRI segmentation. *Sensors.* 2021;21:3363. doi:10.3390/s21103363

54. Tong Q, Li C, Si W, et al. RIANet: recurrent interleaved attention network for cardiac MRI segmentation. *Comput Biol Med.* 2019;109:290-302. doi:10.1016/j.compbiomed.2019.04.042

55. Zhang J, Jiang Z, Dong J, Hou Y, Liu B. Attention gate ResU-Net for automatic MRI brain tumor segmentation. *IEEE Access.* 2020;8:58533-58545. doi:10.1109/ACCESS.2020.2983075

56. Küstner T, Hepp T, Fischer M, et al. Fully automated and standardized segmentation of adipose tissue compartments via deep learning in 3D whole-body MRI of epidemiologic cohort studies. *Radiol Artif Intell.* 2020;2:e200010. doi:10.1148/ryai.2020200010

57. McBride A, Khan HI, Aitken D, et al. Does cartilage volume measurement or radiographic osteoarthritis at baseline independently predict ten-year cartilage volume loss? *BMJ Musculoskelet Disord.* 2016;17:54. doi:10.1136/s12891-016-0900-7

58. Gilbert S, Moore D, Casey J, Crisco J. Quantification of carpal cartilage facet morphology using Micro-CT. In: 55th annual meeting of the orthopaedic research society. 55th Annual Meeting of Orthopaedic Research Society. 2009. ors.org/transactions/55/1157.pdf

59. Graichen H, Jakob J, von Eisenhart-Rothe R, Englmeier KH, Reiser M, Eckstein F. Validation of cartilage volume and thickness measurements in the human shoulder with quantitative magnetic resonance imaging. *Osteoarthritis Cartilage.* 2003;11:475-482. doi:10.1016/S1063-4584(03)00077-3

60. Woolard JD, Gil AB, Sparto P, et al. Change in knee cartilage volume in individuals completing a therapeutic exercise program for knee osteoarthritis. *J Orthopaedic Sports Phys Therapy.* 2011;41:708-722. doi:10.2519/JOSPT.2011.3633

61. Gemme L, Nardotto S, Dellepiane SG. Automatic MPST-cut for segmentation of carpal bones from MR volumes. *Comput Biol Med.* 2017;87:335-346. doi:10.1016/J.COMPBIO.2017.06.011

62. Raj A, Vishwanathan S, Ajani B, Krishnan K, Agarwal H. Automatic knee cartilage segmentation using fully volumetric convolutional neural networks for evaluation of osteoarthritis. *Proc Int Symp Biomed Imag.* 2018:851-854. doi:10.1109/ISBI.2018.8363705

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**FIGURE S1** Schematic structure of an attention layer. Feature maps from encoder and decoder propagate through 1 x 1 size aligning convolutions and are added up
added so that the resulting feature map values are modulated according to the feature map values of both sources. After that, feature maps are activated by and normalized by a sigmoid function. Finally, after a bilinear upsampling, attention coefficients are occurred.

**FIGURE S2.** Algorithm for multi-instance splitting of the training data. As the splitting principle is not anatomically justified, we call the obtained instances – pseudo-instances.

**FIGURE S3.** Architecture of the U-Net convolutional neural network

**FIGURE S4.** Architecture of the truncated U-Net convolutional neural network.

**FIGURE S5.** Architecture of the U-Net convolutional neural network enriched with attention layers (AL). The attention layers were built into skip connections allowing to use information from the encoder path and to filter it eventually emphasizing important regions by combining low- and high-level information from both paths.

**FIGURE S6.** Architecture of the truncated U-Net convolutional neural network enriched with attention layers (AL). The attention layers were built into skip connections allowing to use information from the encoder path and to filter it eventually emphasizing important regions by combining low- and high-level information from both paths.

---

**How to cite this article:** Vladimirov N, Brui E, Levchuk A, et al. CNN-based fully automatic wrist cartilage volume quantification in MR images: A comparative analysis between different CNN architectures. *Magn Reson Med.* 2023;90:737-751. doi: 10.1002/mrm.29671