Computers in Biology and Medicine 148 (2022) 105932

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Ensemble learning and test-time augmentation for the segmentation of mineralized cartilage versus bone in high-resolution microCT images

Jean Léger [a,*], Lisa Leyssens [b,c], Greet Kerckhofs [b,c,d,e], Christophe De Vleeschouwer [a,**]

[a] ICTEAM, UCLouvain, Louvain-la-Neuve 1348, Belgium
[b] iMMC, UCLouvain, Louvain-la-Neuve 1348, Belgium
[c] Institute of Experimental and Clinical Research, UCLouvain, Louvain-la-Neuve 1348, Belgium
[d] Department of Materials Science and Engineering, KU Leuven, Leuven 3000, Belgium
[e] Prometheus, Division of Skeletal Tissue Engineering, KU Leuven, Leuven 3000, Belgium

ARTICLE INFO

Keywords:
Cartilage versus bone
CNN
Ensembling
High-resolution microCT
Mineralized tissue segmentation
TTA
Uncertainty

ABSTRACT

High-resolution non-destructive 3D microCT imaging allows the visualization and structural characterization of mineralized cartilage and bone. Deriving statistically relevant quantitative structural information about these tissues, however, requires automated segmentation procedures, mainly because manual contouring is user-biased and time-consuming. Despite the increased spatial resolution in microCT 3D volumes, automatic segmentation of mineralized cartilage versus bone remains non-trivial since they have similar grayscale values. Our work investigates how reliable 2D segmentation masks can be predicted automatically based on a (set of) convolutional neural network(s) trained with a limited number of manually annotated samples. To do that, we compared different strategies to select the 2D samples to annotate and considered ensemble learning and test-time augmentation (TTA) to mitigate the limited accuracy and robustness resulting from the small number of annotated training samples. We show that, for a fixed amount of annotated image samples, 2D microCT slices to annotate should preferably be selected in distinct 3D volumes, at regular intervals, rather than being grouped in adjacent slices of a same 3D volume. Two main lessons are drawn regarding the use of ensembles or TTA instead of a single model. First, ensemble learning is shown to improve segmentation accuracy and to reduce the mean and standard deviation of the absolute errors in cartilage characteristics obtained with different initializations of the neural network training process. In contrast, TTA appears to be unable to improve the model's robustness to unlucky initializations. Second, both TTA and ensembling improved the model's confidence in its predictions and segmentation failure detection.

1. Introduction

Microfocus X-ray computed tomography (microCT) is a 3D X-ray-based imaging technique that allows the visualization of the internal structures of mineralized tissues thanks to its high spatial resolution (<1 μm voxel size). In contrast to histological sectioning, microCT is non-destructive, which allows the 3D structural information of the imaged tissue to be captured without a limit on depth resolution. MicroCT is accepted as a standard tool for structural characterization of mineralized tissue, such as bone. Interestingly, advances in the achievable resolution nowadays make it possible to distinguish mineralized cartilage from bone. This comes from the fact that osteocyte and chondrocyte lacunae in bone and cartilage, respectively, can be visualized using microCT [1,2]. Such visualization capabilities pave the way to numerous applications relying on the 3D structural characterization of not only bone, but also mineralized cartilage. For instance, in the bone-to-tendon interface, the attachment between bone and mineralized fibrocartilage is a common site of injury due to aging, sports, or disease such as diabetes. Unfortunately, a rupture is very complex to repair because the native properties of the attachment site are not recovered [3]. Using microCT, the native structural properties of the interface could be assessed and used to design biomimetic scaffolds for improved artificial tissue regeneration. Alternatively, microCT can be used to better understand the changes in the thickness of mineralized cartilage in the bone-to-cartilage interface that occur in osteoarthritis and other joint diseases and, in turn, better understand the effect of the disease on the proper function of the interface [4]. Finally, in bone tissue engineering applications, microCT allows the assessment of endochondral bone formation (i.e., the fraction of newly formed bone

* Correspondence to: ELEN, Place du Levant 3/L5.03.02, 1348 Louvain-la-Neuve, Belgium.
** Corresponding author.
E-mail addresses: jean.leger@uclouvain.be (J. Léger), christophe.devleeschouwer@uclouvain.be (C.D. Vleeschouwer).

https://doi.org/10.1016/j.compbiomed.2022.105932
Received 30 March 2022; Received in revised form 6 July 2022; Accepted 30 July 2022
Available online 2 August 2022

versus mineralized cartilage). This is important to assess maturity of tissue engineered constructs and the effect of the construct ingredients such as growth factors and cell types on the maturity [5].

For these applications, it is important to be able to assess the 2D and 3D structural characteristics of mineralized cartilage, such as volume, thickness, porosity, and number of chondrocyte lacunae [4,6,7]. Importantly, the evaluation of these parameters relies on the preliminary segmentation of mineralized cartilage versus bone in high-resolution microCT images. However, segmenting mineralized cartilage versus bone remains challenging as the grayscale density attenuation of bone and mineralized cartilage is very similar because of the minor difference in mineralization degree between subchondral bone and mineralized cartilage [4], as well as their highly irregular interface [8]. Mineralized cartilage and bone differ rather in texture because of the presence of chondrocyte lacunae in cartilage and no or limited lacunae in the subchondral bone. Because of the difficulty of separating mineralized cartilage from bone, the two tissues are still mostly delineated manually, which is time consuming and user-biased. Also, we expect manual segmentation to become more challenging in the future with the tremendous increase in data size [9]. Hence, advanced automatic segmentation tools for mineralized cartilage versus bone are required.

Various neural architectures have been considered to automate image segmentation [10,11]. Deep learning-based approaches, such as U-Net [12], have become state-of-the-art for automatic biomedical image segmentation [13] and their applicability to unmineralized cartilage segmentation has already been shown [9]. Most previous work on unmineralized cartilage segmentation focused on magnetic resonance (MR) images, whereas segmentation on microCT-based datasets is less common [4,14,15]. The advantages of using microCT compared with MR imaging is that the acquisition time is much lower [16] and the achievable spatial resolution is much higher (i.e., <1 μm and 20 μm for microCT and MR imaging, respectively) [17]. For the imaging of mineralized tissue, as is the case in this work, microCT is the natural choice since no contrast-enhancing techniques are needed. As far as we know, despite the popularity of CNN to segment tissues and organs in CT images [18], automatic mineralized cartilage segmentation in microCT images using convolutional neural networks (CNNs) has been studied in [4] and our preliminary work [15] only. Both papers use U-Net as a CNN architecture because this model can learn complex discriminant features, such as textures, from data automatically. Hence, it is suited for capturing the characteristic texture of mineralized cartilage observed due to the presence of chondrocyte lacunae. Also, the approach is fully automatic and fast, thereby increasing the segmentation speed a lot compared with manual segmentation.

However, despite their high segmentation accuracy, deep learning-based approaches still have several known limitations, especially for (bio)medical image segmentation. First, fully supervised approaches suffer from the limited amount and accuracy of labeled data, since manual annotations require expert knowledge and may contain errors due to the smooth, progressive transition between distinct zones [19]. This is particularly true for mineralized cartilage versus bone segmentation, since the number of samples is typically limited by the number of animals available and the number of annotations is further limited by the availability of experts. Also, the composition of a representative training set is challenged by the high irregularity of mineralized cartilage's shapes and borders. Second, similarly to different annotators that may have different annotation styles, different CNN initializations explore different modes (i.e., learned functions), which can lead to significantly different segmentation outputs [20]. Third, CNNs trained for segmentation are poorly calibrated, meaning that they are not able to estimate the confidence in their predictions correctly [21]. Hence, they cannot detect their failure automatically. This last limitation is even more problematic for small datasets, since errors are more likely to occur in this case. Also, considering the difficulty of distinguishing mineralized cartilage from subchondral bone even for an expert due to the very minor difference in mineralization between the two tissues, we believe that obtaining a reliable confidence metric for automatic segmentation is valuable for potential downstream tasks or visual inspection of the automatic segmentation results.

In this work, we analyzed how these limitations affect mineralized cartilage versus bone segmentation in microCT images and thoroughly investigated how to address these limitations. More precisely, we considered using multiple CNN predictions to improve segmentation accuracy and associate with every pixel a score that better reflects the probability that the pixel belongs to mineralized cartilage. Also, we evaluated the impact of multiple CNN predictions on the evaluation of mineralized cartilage area, porosity, and number of chondrocyte lacunae in 2D images. Since we are interested in situations with scarce data, we further evaluated the impact of varying the number of training samples (and the strategy to choose them) on the segmentation accuracy and confidence estimation.

Our study resulted in three main recommendations regarding the design of CNN-based automatic solutions for microCT-based image analysis:

- First, we demonstrated that, for given annotation resources, annotating a small set of 2D slices regularly sampled in distinct 3D microCT volumes resulted in models that generalized better than the ones trained from entirely annotated 3D microCT volumes.
- Second, it revealed that the accuracy of the prediction made on unseen data and its robustness regarding model initialization were improved more by combining models trained with distinct initializations than by applying the same model on multiple transformed versions of the input.
- Third, input transformations and different initializations appeared to be both effective in estimating the model's confidence in its predictions. Also, model prediction failures could be located from the inconsistency, and thus the entropy, of the multiple predictions associated with transformed versions of the input or with an ensemble of models obtained with different initializations.

2. Related work

Our work considers multiple CNN predictions to improve segmentation accuracy and confidence estimation. To do so, we leverage methods that are used for uncertainty estimation in deep neural networks. In the following section, we introduce the methods that have been proposed to estimate prediction uncertainty in networks and discuss the calibration of this uncertainty. We then consider previous studies involving deep learning for cartilage segmentation and we present lacunae in mineralized cartilage in detail.

2.1. Uncertainty estimation with deep neural networks

Bayesian neural networks provide a mathematical formalism for estimating the uncertainty in a model's predictions due to a lack of knowledge of the model (i.e., epistemic uncertainty). However, this approach is computationally intractable for non-trivial cases. Approximations were referenced in [22] and include Laplace approximation [23], Markov chain Monte Carlo methods [24], variational Bayesian methods [25,26], and Monte Carlo Batch Normalization [27]. Alternatively, Monte Carlo sampling can be used to generate a set of predictions and estimate uncertainty by aggregating the predictions. The set of predictions can be obtained from an ensemble of models trained separately with different initializations [22] or from dropout at test time [28]. The major drawback of deep ensembles in comparison to test time dropout (TTD) is that they require training several models and storing all of them. However, deep ensembles trained with random initializations have been shown to produce a more diverse set of predictions than subspace sampling methods such as weight averaging, TTD, and various versions of local Gaussian approximations [20]. This results in improved uncertainty estimation. Nevertheless, TTD remains

very popular, since it is easy to use and has no memory overhead compared with a single model. Fewer approaches have been proposed for estimating uncertainty due to the lack of informativeness provided by the data on the scene to interpret (i.e., aleatoric uncertainty). A mapping from input data to aleatoric uncertainty was learned using a Bayesian deep learning framework [29]. However, this method is not suitable for already trained models as it requires adapting the loss function and network architecture. Alternatively, data augmentation at test time (TTA) was used to generate several predictions per test sample and subsequently estimate the aleatoric uncertainty [30,31]. In this work, model ensembling and TTA were selected as uncertainty estimation methods, whatever the source of uncertainty (epistemic or aleatoric).

2.2. Confidence calibration with deep neural networks

Modern (deep) neural networks have been shown to be poorly calibrated [21]. This means that the likelihood estimates they produce are not representative of the true probability of correctness. Several methods have been proposed to improve the calibration of single deep neural networks. These include using a proper scoring rule as the loss function (e.g., cross-entropy loss instead of Dice loss) [21,22], using weight decay, reducing the model capacity, and avoiding batch normalization [21]. In order to improve the calibration further, some recent approaches have considered post hoc calibration methods for segmentation [32,33]. However, little attention has been paid to uncertainty as a method for improving the calibration of deep neural networks in a segmentation context. A first investigation in that respect was published in [34], where ensembling was proposed as a confidence calibration approach for fully convolutional neural networks (FCNs). The ability of calibrated FCNs to predict the segmentation quality of structures and detect out-of-distribution test examples was then investigated. Also, a systematic comparison of Bayesian and non-Bayesian approaches (Monte Carlo dropout and ensembles) was performed based on segmentation accuracy, probability calibration, and uncertainty on out-of-distribution datasets [35]. However, using uncertainty as a strategy to improve model prediction calibration has not been studied in the context of quantitative analysis of biomedical images, where the purpose is to reduce the standard deviation of tissue structural parameters estimation rather than to detect out-of-distribution samples in the test set.

2.3. Cartilage segmentation with deep learning

Unmineralized knee cartilage was segmented on MR images using a triplanar CNN [36], Segnet [37], 2D U-Net [38], 2D U-Net complemented with conditional random fields to improve 3D consistency [39] and 3D approaches [40,41]. In [42], an ensemble-based learning approach was used for cartilage segmentation on knee MR images of patients with osteoarthritis. In [43], Monte Carlo dropout was used to improve the segmentation of unmineralized femoral cartilage in ultrasound images. Fewer approaches consider automatic unmineralized cartilage segmentation on CT images. They include an atlas-based segmentation of acetabular cartilage [44] and a registration-based segmentation of knee articular cartilage [45] in contrast-enhanced clinical CT images. In [14], craniofacial cartilage was segmented in 3D microCT volumes of embryonic mice stained with phosphotungstic acid, using extremely sparse annotations (e.g., annotating only a few selected slices in a volume). An ensemble-based approach was used to generate pseudo-labels on unlabeled data, as well as an estimation of their uncertainty. The uncertainty information was then used to guide a self-training process. However, none of those studies considered the automatic segmentation of mineralized cartilage with deep learning. Our previous paper considered the segmentation of mineralized cartilage and bone in high-resolution microCT images [15]. Recently, mineralized cartilage was segmented with a deep learning model in microCT images [4]. This work focuses on the segmentation of mineralized cartilage in the knee from twelve New Zealand White rabbits and further assesses the thickness of the mineralized cartilage. However, the latter two approaches do not discuss the model's confidence in its predictions or the limitations of their methods when a low amount of annotated data is available - a common case in practice that lies at the core of our study.

2.4. Lacunae in mineralized cartilage

A specificity of this work is to focus on mineralized cartilage. Mineralized cartilage occurs at three types of interfaces or transitions: the interface between bone and fibrocartilage (in the bone–tendon interface or enthesis) [3,46], the osteochondral interface (i.e., the bone–cartilage interface) [4], and during endochondral bone formation [5]. Mineralized cartilage is characterized by the presence of lacunae that contain the cartilage cells or chondrocytes. Human chondrocytes occur singly, in pairs, or as clusters within a lacuna, with a cluster comprising three or more chondrocytes within the enclosed lacunar space [47]. These lacunae can be observed on microCT images [1,2] and can be identified thanks to their difference in texture compared with bone. The main challenge for automatic segmentation of mineralized cartilage is the absence of contrast difference between cartilage and bone and the irregularity of the border between the two tissues, making the annotation ambiguous. This is the main motivation for using advanced segmentation tools, such as deep learning-based approaches, coupled with mechanisms to associate a probability, also named confidence, with its predictions.

3. Materials and methods

3.1. Data acquisition

3D high-resolution microCT volumes of 12 murine bone-to-Achilles tendon interfaces were considered in this study. After harvesting, all the samples were fixed in 4% formaldehyde for 16 h and then stored at 4 °C in a phosphate-buffered saline solution (PBS) until further processing. Next, the samples were further dissected to isolate the tendon and bone and to remove all tissue that was unnecessary for the experiments. A Phoenix Nanotom M - Computed tomography system (GE Sensing & Inspection Technologies GmbH, Wunstorf, Germany) was used with the following scanning parameters: 60 kV, 87 μA, 1.25 μm voxel size, 2400 images, 500 ms exposure time, 20 min scan time. The microCT volumes were reconstructed with the Datos|x software (GE Measurement and Control Solutions) and exported as XY slices (.tiff). A beam hardening correction of eight was applied and three different filters were used: inline media, ROI-CT, and filter volume. The samples were left over tissues from an experiment approved by the Ethics Committee of the University of Leuven (P101/2014) and in compliance with the ARRIVE guidelines (https://arriveguidelines.org/).

3.2. Data preprocessing

The mineralized cartilage was manually annotated on all the slices of the microCT images by an expert using CTAn (Bruker MicroCT, Kontich, Belgium). These annotations served as the ground-truth in this work and were stored as 12 3D binary masks with the same size and resolution as the 3D microCT volumes. The 3D volumes and binary masks were cropped such that every slice had a size of 1024 × 1024 pixels. The cropping operation preserved the entire region of interest (i.e., mineralized cartilage and bone) and was not part of the data augmentation process. The images were normalized based on the mean μ_{tr} and standard deviation σ_{tr} of the pixels computed on the training set. As we propose a 2D segmentation approach, we define a stack of images as a set of 2D images belonging to the same 3D volume.

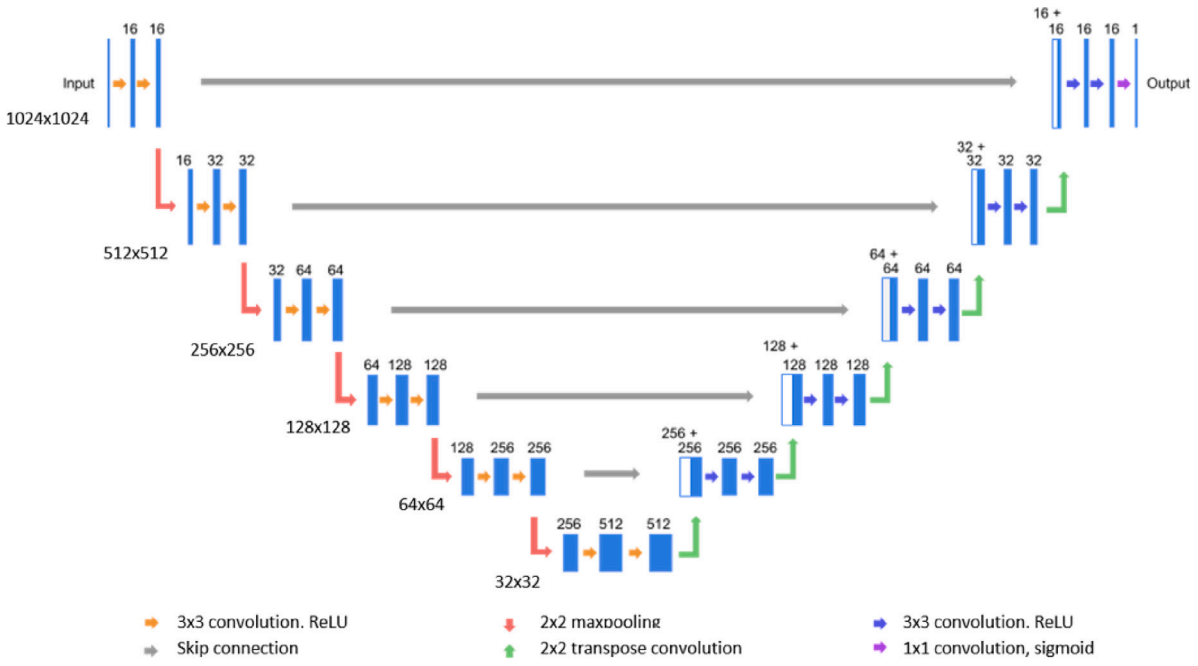

Fig. 1. U-Net architecture. The downsampling part is composed of convolutional layers with kernels of spatial size 3 × 3 (3 × 3 convolution), rectified linear unit activations (ReLU), and maxpooling layers computed on an area of 2 × 2 pixels (2 × 2 maxpooling). The number of feature maps in the first layers is set to 16 and increases as the feature map resolution decreases. In the upsampling part, the maxpooling layers are replaced by transpose convolutional layers with kernels of spatial size 2 × 2 (2 × 2 transpose convolution). Skip connections concatenate feature maps of the downsampling with feature maps of the upsampling part. A convolution with a kernel of spatial size 1 × 1 (1 × 1 convolution) generates a single feature map at the output of the network, which is passed through a sigmoid activation function to obtain the final prediction.

3.3. Convolutional neural networks

Cartilage was segmented on microCT images using the 2D U-Net fully convolutional neural network [12,48]. As depicted in Fig. 1, U-Net is an encoder–decoder architecture with skip connections. Such architecture makes it possible to capture features at multiple resolutions while maintaining the high resolution in the output. In encoder–decoder architectures, the encoder gradually compresses the input image into a low-resolution latent space representation. This representation captures the semantic information of the input that is useful for the segmentation task. The decoder aims to predict the high-resolution output from the low-resolution latent space representation. The drawback of such encoder–decoder architectures is that the input image spatial resolution is barely recovered using only the spatially low-resolution latent space representation. Hence, U-Net uses skip connections between the encoder and the decoder such that feature maps from the downsampling part of the network are copied to the upsampling part. Using both copied and upsampled feature maps in the decoder allows leveraging both high-resolution and contextual pattern information.

More precisely, the encoder is composed of convolutional and pooling layers. Convolutional layers compute the cross-correlation between the layer input (i.e., the input image or the output of a previous layer) and a set of kernels, also named filters, followed by an activation function, in order to extract features. Thus, the output of a convolutional layer is a set of feature maps, where each feature map is generated with a different kernel. The kth pre-activation $Z_k \in \mathbb{R}^{n_i \times n_j}$ is obtained as the cross-correlation between the padded input $I \in \mathbb{R}^{(n_i+2p)\times(n_j+2p)\times n_c}$ and kernel $H_k \in \mathbb{R}^{n_m \times n_m \times n_c}$ (with a bias term b_k), where n_i, n_j, n_m are spatial dimensions, n_c is the channel dimension, and $2p$ is the size of the padding along each spatial dimension of I. The goal of the padding is to preserve the spatial dimensions between the input and the pre-activation. The padding consists in concatenating p zeros at the beginning and the end of I along each spatial axis. The size of the padding is chosen such that $n_m - 1 = 2p$. Formally, the cross-correlation using input padding is written as

$$Z_k(i,j) = \sum_{m=0}^{n_m-1} \sum_{n=0}^{n_n-1} \sum_{c=0}^{n_c-1} I(i+m, j+n, c) H_k(m,n,c) + b_k, \quad (1)$$

for $0 \le i < n_i$ and $0 \le j < n_j$. An element-wise activation $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is then applied to the pre-activation to obtain the kth feature map $A_k \in \mathbb{R}^{n_i \times n_j}$. Formally, the activation is written as

$$A_k(i,j) = \sigma(Z_k(i,j)). \quad (2)$$

Pooling layers replace a small neighborhood of a feature map with some statistical information (mean, maximum, etc.) about the neighborhood, therefore reducing spatial resolution. The neighborhood is generally a region of 2 × 2 features. The decoder also relies on convolutional layers followed by an activation, whereas pooling layers are replaced by upsampling layers. Upsampling layers insert zeros between the feature map elements, thereby increasing the sampling rate, generally by a factor of two in every spatial dimension. A convolution, generally with kernel spatial size 2 × 2, is then applied to the upsampled feature maps to obtain a dense feature map (i.e., without zeros) at a higher resolution. The combination of the upsampling and convolution operations is also called transpose convolution.

The U-Net encoder stacks the following set of layers several times: two convolutional layers with kernels of spatial size 3 × 3 (in short, 3 × 3 convolutions), each followed by ReLU activations, and a pooling layer computed on a spatial neighborhood of 2 × 2 features. The pooling operation computes the maximum over the considered neighborhood, therefore being called max-pooling (in short, 2 × 2 max-pooling). The number of feature channels is doubled at each downsampling step. The ReLU activation[1] stands for rectified linear

[1] The ReLU is applied element-wise to the feature maps obtained with the convolutional layer. The ReLU activation is fast to compute and has a derivative equal to 1 when $x > 0$, and zero otherwise. Thus, multiplying several ReLU derivatives together when applying the chain rule during the backpropagation algorithm has the property of being either 1 or 0. This helps to reduce the vanishing gradient problem during the backpropagation.

Table 1
Training set configurations and notations as used in the rest of this work. A stack is defined as a set of 2D images from the same 3D volume.

	Config 1	Config 2	Config 3	Config 4	Config 5
Number of training stacks, a	1	2	2	8	8
Number of images per stack, b	10	5	100	5	100
Notation, C(a, b)	C(1,10)	C(2,5)	C(2,100)	C(8,5)	C(8,100)

unit activation [49] and is defined by $\sigma_{ReLU}(x) = \max(0, x)$. The bottleneck of the network is composed of two 3×3 convolutions, each followed by ReLU activations. Every step in the decoder consists of transpose convolution operations with kernels of spatial size 2×2 (in short, 2×2 transpose convolutions), halving the number of features, a concatenation with the correspondingly feature maps from the encoder, and two 3×3 convolutions, each followed by a ReLU activation. The output layer consists of a 1×1 convolution, followed by a sigmoid activation. The 1×1 convolution operates only on the channel dimension of the feature maps to produce a single-channel output. The sigmoid activation is defined by

$$\sigma_{sig}(Z(i,j)) = \frac{1}{1 + \exp(-Z(i,j))}, \tag{3}$$

where $Z(i, j)$ is the output of the 1×1 convolution.

3.4. Learning strategy

3.4.1. Base learner architecture

2D U-Net was used as a base learner in our work. The 2D U-Net architecture was preferred to its 3D counterpart since it reduces the graphical processing unit (GPU) running time associated with this study greatly. In practice, 2D and 3D U-Net showed similar segmentation performance in terms of the Dice similarity coefficient (DSC) for mineralized versus bone segmentation in microCT images, with slightly better 3D consistency with the 3D approach [15]. However, 2D U-Net enables one to work at full resolution on 2D slices directly whereas a 3D approach would require extracting sub-volumes for model training and predicting on overlapping sub-volumes at inference. Also, reducing the number of slices is straightforward to study in 2D, since every training sample is one slice. Finally, a 2D approach consumes less storage space, which becomes critical when model ensembles are used.

The training set (i.e., the set of images that are used during the learning process in order to automatically tune the kernel parameters to perform the segmentation) was denoted by $D_{tr} = \{(x_j, y_j)\}_{j=0}^{J-1}$, where x_j and y_j denote the jth microCT slice and its corresponding ground-truth binary segmentation mask, respectively, and J is the total number of slices in the training set. The final activation function of the network is a sigmoid, such that the model learns the posterior $p(y|x, \theta)$, where θ are the inner weights of the model. The most probable class label was assigned to each pixel individually according to the following decision rule

$$\hat{y}(i) = \underset{c(i)\in\{0,1\}}{\arg\max}\, p(Y(i) = c(i)|x, \theta),$$

where i is the pixel index within the test set. The number of resolutions of U-Net was set to six,[2] such that the receptive field of the model in the original resolution reached 284×284 pixels. The batch size was set to five, which was the maximum size affordable on our 11 Gb graphical processing units (GPU). The network was trained with the Dice loss. The Adam optimization algorithm was used with a fixed learning rate of 10^{-4}. The source code is publicly available on https://github.com/legerjean/cartilage_segmentation.

[2] Four resolutions were tested but led to unrealistic disconnected cartilage predictions, probably due to lack of context information. Eight resolutions allowed the receptive field to cover the full slice as recommended in [13] but did not show enhanced performance compared with six resolutions.

3.4.2. Training and test set configurations

Cartilage segmentation and structural characterization were performed using different training set configurations, the goal being to build a training set efficiently with a limited amount of annotations. Five training set configurations were considered, each of them using a different number of training images, and/or sampling the images from a different number of stacks, as shown in Table 1. The notation $C(a, b)$ corresponds to the selection of b images in a stacks, for a total of ab images. The images were sampled uniformly within the stacks. Slices from four other 3D volumes were used as the test set. As only the number of resolutions of U-Net had been tuned on the test data, and to keep data available for training and testing, no validation set was considered here. Cross-validation was not considered in this work due to the lack of computational resources for training the large number of models required for ensemble strategies.

3.4.3. Data augmentation during training

The corpus of training data was augmented according to the following plausible transformations. Rotations were applied between 0° and 360°. The image brightness was varied using a random offset added to the normalized image, such that $x \leftarrow x + u\mathcal{I}$, where $u \sim U[-\tau, \tau]$ and $\mathcal{I} \in \mathbb{R}^{1024\times1024}$ is a matrix with all entries equal to 1. τ was set to $40/\mu_{tr}$ as the range of variation of the slice-wise average cartilage intensity reached 40 in the full training set (i.e., considering the eight 3D volumes) before normalization. Finally, additive white Gaussian noise with standard deviation σ_{bg}/μ_{tr} was applied to each pixel of the image. σ_{bg} is the standard deviation of the background noise measured in the training images before normalization and reached 15. A different transformation was applied for each training slice and each epoch. The number of epochs n_{epochs} was adapted as a function of the number of slices in the training set, n_{slices}, following $n_{epochs} = n_{max}/n_{slices}$, where n_{max} is set to 90 000 to reach convergence.

3.4.4. Inference

At inference, additive white Gaussian noise was applied on the input images, as for the training phase. Indeed, adding noise to the input was seen to improve segmentation accuracy. Multiple predictions for a given test sample were obtained either through a Monte Carlo approach at training using different model initializations or at inference using test-time augmentation (TTA). The first strategy resulted in K different models, each of them providing one prediction. The second strategy resulted in a single model for which K different predictions were computed. The transformations that were used during TTA were the same as during the training phase. Both strategies generated a set of predictions $\hat{\mathcal{Y}}(i) = \{\hat{y}_0(i), \hat{y}_1(i), \ldots, \hat{y}_{K-1}(i)\}$, where i is the pixel index within the test set. These diverse predictions were further used to improve cartilage segmentation, calibrate the model, detect failure, or improve cartilage characterization evaluation. An example of predictions obtained with different model initializations is shown in Fig. 2.

3.5. Metrics

3.5.1. Model accuracy

The Dice similarity coefficient (DSC) was used to assess the quality of the automatic segmentation of mineralized cartilage in comparing the binary masks derived from the manual annotations and the automatic segmentation. The DSC measures the overlap between two binary

Table 2
Automatic segmentation performance for different training set configurations (one per line) using ensembling and TTA. Twenty different segmentation outputs were generated using different initializations. The mean, standard deviation (SD), and worst and best cases of the DSC (averaged across all test slices) were computed across the twenty outputs. Majority voting (MV) was performed pixel-wise on the twenty predictions and the DSC was computed on the fused output. Then, TTA was performed using the models with the worst and best initializations. The mean and standard deviation were computed from the twenty averaged (over test slices) DSCs obtained with twenty different augmentations. Majority voting was performed between the predictions obtained with different augmentations and the DSC was computed on the fused output.

Training set	Different initializations				TTA for worst initialization		TTA for best initialization	
configuration	Mean ± SD	Worst case	Best case	MV	Mean ± SD	MV	Mean ± SD	MV
C(1,10)	0.866 ± 0.009	0.844	0.877	0.874	0.835 ± 0.011	0.837	0.873 ± 0.004	0.877
C(2,5)	0.880 ± 0.006	0.871	0.889	0.889	0.869 ± 0.002	0.872	0.889 ± 0.002	0.892
C(2,100)	0.883 ± 0.010	0.845	0.899	0.893	0.855 ± 0.012	0.859	0.892 ± 0.010	0.898
C(8,5)	0.907 ± 0.002	0.902	0.912	0.912	0.902 ± 0.001	0.905	0.910 ± 0.002	0.913
C(8,100)	0.916 ± 0.006	0.900	0.923	0.923	0.907 ± 0.006	0.916	0.921 ± 0.002	0.924

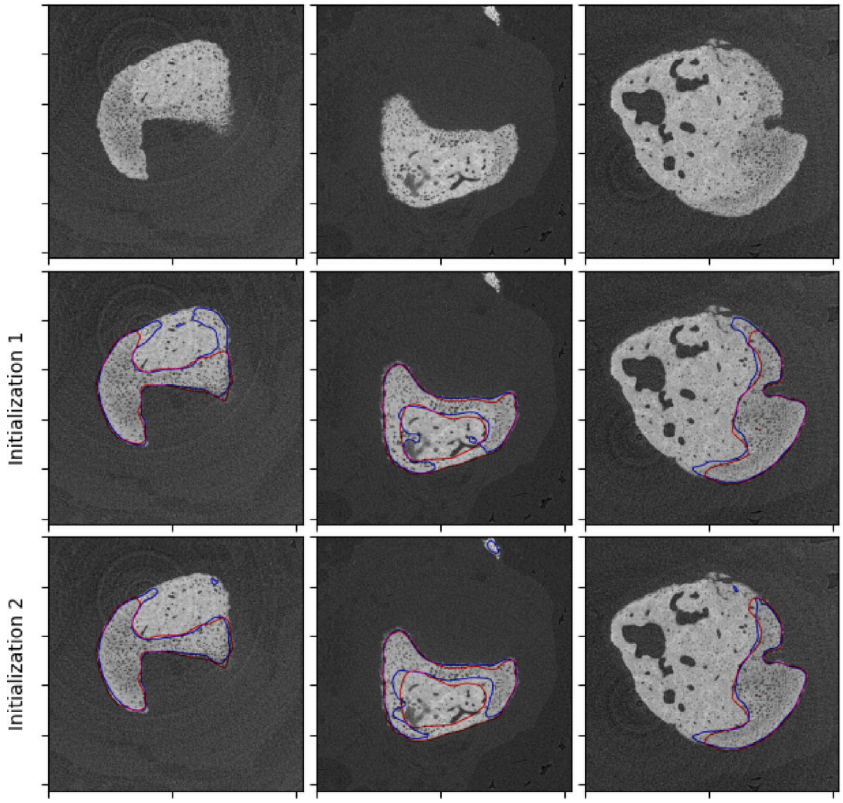

Fig. 2. Manual and automatic segmentation of mineralized cartilage on three high-resolution microCT slices. The automatic segmentation method uses a single CNN trained on 800 annotated slices. The automatic segmentation output is shown for two different initializations. The CNN is always highly confident in its predictions, even in regions where another CNN trained with a different initialization or the manual segmentation disagrees. Red: manual segmentation, blue: automatic segmentation.

masks. Formally, if two binary masks are denoted by A and B, the DSC is defined by

$$\mathrm{DSC}(A;B) = \frac{2|A \cap B|}{|A| + |B|}. \tag{4}$$

The cartilage structure was characterized in 2D using three metrics: its area, the number of chondrocyte lacunae (also denoted as closed pores) within it, and its porosity. These metrics were used to compare the manual and automatic segmentation approaches further. The cartilage area is the sum of the annotated or automatically segmented pixels multiplied by the pixel area of 1.25^2 μm^2. The chondrocyte lacunae were automatically segmented from the microCT images using the Otsu algorithm [50]. This algorithm separates pixels into two classes (i.e., chondrocyte lacunae and surrounding mineralized cartilage) based on their intensity using a single threshold. It is suited for images having a bimodal intensity histogram. This is the case here, since chondrocyte lacunae are significantly darker than the surrounding mineralized cartilage. The optimal threshold $\hat{\tau}$ is such that it minimizes the intra-class intensity variance. Formally, this can be expressed as

$$\hat{\tau} = \arg\min_{\tau} \left[\omega_0(\tau)\sigma_0^2(\tau) + \omega_1(\tau)\sigma_1^2(\tau)\right], \tag{5}$$

where the weights ω_0 and ω_1 are the probabilities of the two classes separated by the threshold τ and σ_0 and σ_1 are variances of these two classes. Intuitively, the optimal threshold corresponds to an intensity located between the two modes of the histogram. The porosity is defined as the ratio between the areas of the chondrocyte lacunae and the mineralized cartilage.

3.5.2. Model calibration

Considering a binary segmentation task, let $\hat{p}_i$ be the estimated probability for a pixel to belong to the positive class, as obtained from the deep learning approach. Ideally, the probability estimate $\hat{p}_i$ should be calibrated, which means that $\hat{p}_i$ should reveal the true probability of the considered pixel's belonging to the positive class. When a single prediction was available, we used the output of the network (i.e., the output of a sigmoid layer) as a natural choice to obtain $\hat{p}_i$. However, this output is known to be uncalibrated for CNNs trained with Dice

loss and batch normalization [21]. Therefore, we used the set $\hat{\mathcal{Y}}(i)$ of predictions produced during testing to improve model calibration, as proposed in [34]. Following the law of large numbers, the probability is approximated by the frequency of realizations. Hence, when multiple predictions were available, we computed the estimated probability $\hat{p}_i$ as the frequency of cartilage predictions in $\hat{\mathcal{Y}}(i)$. In this case, $\hat{p}_i$ was obtained by

$$\hat{p}_i = \frac{1}{K}\sum_{k=0}^{K-1} \hat{y}_k(i). \tag{6}$$

We used the expected calibration error (ECE) as summary statistics for calibration assessment. Since we are dealing with a binary classification problem, we adopted the calibration error metric introduced in [51,52]. Interested readers may refer to [21] for a generalization of this metric to problems involving more than two classes. To compute the ECE, the pixel-wise cartilage estimated probability maps $\hat{p}_i$ were firstly sorted and partitioned into $Q + 1$ bins, where the qth bin contained probabilities in the range $I_q = \left]\frac{q-1}{Q}, \frac{q}{Q}\right]$ and $q \in [0, 1, \ldots, Q]$. We chose $Q = 20$, since ensembling and TTA considered 20 different predictions when model calibration was assessed. The ECE is defined as:

$$ECE = \sum_{q=0}^{Q} w_q \left| f_q - \bar{p}_q \right|, \tag{7}$$

where w_q is the empirical fraction of all pixels whose estimated probability falls into bin q, f_p is the true fraction of cartilage pixels in bin q, and $\bar{p}_q$ is the mean of the estimated probabilities for the instances in bin q. The lower the value of the ECE, the better the model calibration will be.

3.5.3. Uncertainty estimation

Under the assumption that the estimated probability map $\hat{p}_i$ is calibrated, we were interested in determining whether it could be used to locate regions where the model was highly uncertain, and therefore prone to segmentation errors. A calibrated estimated probability map will reach 0.5 in regions where the model is highly uncertain and 0 or 1 in regions where the model is highly confident. However, uncertainty is best captured with a metric that is high for uncertain predictions and low for confident predictions. This is provided by entropy [31]. Pixel-wise entropy can be approximated as

$$H\left(Y(i)|X\right) \approx -\hat{p}_i \log_2\left[\hat{p}_i + \epsilon\right] - \left(1 - \hat{p}_i\right)\log_2\left[1 - \hat{p}_i + \epsilon\right], \tag{8}$$

where $Y(i)$ and X are the pixel-wise actual label and input image random variables, respectively, and $\epsilon = 10^{-10}$ allows approximating the entropy for $\hat{p}_i$ equal to 0 or 1. Besides pixel-wise uncertainty, subject-wise uncertainty estimations are needed to quantify the reliability of subject segmentation performance and possibly detect model failure [53]. Therefore, we defined a metric that computed the slice-wise uncertainty level as

$$s_H = \frac{100}{N_m}\sum_{i \in S_m} H\left(Y(i)|X\right), \tag{9}$$

where N_m is the number of pixels such that $\hat{p}_i > 0.5$ in slice m, and S_m is the set of indices of pixels belonging to slice m. Normalization by N_m (i.e., the predicted slice-wise cartilage area, as obtained with a majority voting approach between the multiple predictions) yields an uncertainty metric that is as independent as possible of the cartilage area. This definition is close to [34] but adapted for binary classification.

4. Results

4.1. Model accuracy

4.1.1. Segmentation performance

The goals of the first sets of experiments were (i) to evaluate the impact of the number of stacks and images in the training set on segmentation performance, and (ii) to study how ensembling (considering twenty initializations) and TTA (considering twenty augmented versions of the input) improve the accuracy and robustness against unlucky initializations for cartilage segmentation. DSC results for different training configurations are provided in Table 2. The reported DSC results are always averaged DSC results computed across the 290 test slices. For each configuration, twenty models were trained with different initializations. The mean, standard deviation, and worst and best cases for the averaged DSC results were computed across the twenty initializations. The twenty segmentation results were also fused pixel-wise using majority voting and the averaged DSC was computed on the resulting fused output. Formally, majority voting classifies the ith pixel as cartilage if $\hat{p}_i > 0.5$. Also, for each training set configuration, TTA was applied to the two models that produced the best and worst averaged DSC results. For each of these models, twenty predictions were made from a different augmented version of the input. The mean averaged DSC and the averaged DSC associated with a majority vote decision are provided in Table 2.

Table 2 shows that the DSC improves with the number of stacks in the training set. Indeed, DSC reached approximately 0.91, 0.88, and 0.86 when 8, 2, and 1 stacks were considered at training, respectively. Interestingly, the number of training stacks has more impact on the average DSC than the total number of training images. Indeed, model $C(8,5)$ trained with 5 images from 8 different stacks (i.e., 40 images in total) provided better average DSC results than model $C(2,100)$ trained with 100 images from 2 different stacks (i.e., 200 images in total). Similarly, both $C(1,10)$ and $C(2,5)$ were trained on ten images in total but higher DSC results were obtained with model $C(2,5)$, which was trained on images sampled from two separate stacks. We explain this by the fact that neighboring images are highly correlated within a single stack. Hence, diversity in the training set is higher when images are sampled from different stacks or far from each other in a single stack than when many images are annotated in the same stack.

Table 2 also shows that the DSC results reached with majority voting across the predictions obtained from models trained from different initializations (i.e., ensembling) were close to the results reached with the best initialization. Majority voting between the predictions obtained from different test-time augmentations (i.e., TTA) also slightly improved DSC results in comparison to a single model trained with the same initialization. This is the case for both worst and best initializations. However, DSC results obtained after majority voting with the worst initialization were lower than the corresponding results obtained with the best initialization, for all training configurations. This shows that the segmentation obtained with TTA remains largely conditioned by the initialization. Hence, ensembling increases robustness to badly initialized modes, which is not the case for TTA. These results are consistent with the fact that different initializations explore different modes (i.e., learned functions) [20].

4.1.2. Influence on the structural characteristics of the mineralized cartilage

Fig. 3 compares the slice-wise cartilage structural characteristics (i.e., area, number of lacunae, and porosity) computed from the manual segmentation with the distribution of the same characteristics derived from the twenty automatic segmentations obtained from models trained with twenty different initializations. The results are presented for training set configuration $C(8,100)$ and are plotted as a function of the slice index, for one of the test stacks. We observed that the characteristics derived from our $C(8,100)$ deep learning segmentation are close to the ones computed from manual segmentation.

We were, however, interested in analyzing how a smaller training set, as well as the ensembling and TTA inference mechanisms, impacted the cartilage structure and, in particular, their robustness to bad initializations. Therefore, Tables 3 and 4 analyze how the absolute difference between the slice-wise cartilage characteristics computed from the manual and (potentially multiple) deep learning-based cartilage segmentation masks depends on the training configuration. To build this

Table 3
Absolute difference between the cartilage characteristics (i.e. area, number of lacunae and porosity) measured for the automatic and manual segmentation approaches. 25 single models have been trained with different initializations. 5 ensembles have been built, each of them considering 5 single models. TTA was applied on every single model, considering twenty augmented versions of the input. Cartilage characteristics are computed slice-wise. Ensembling and TTA consider a fusion at the level of the characteristics (by averaging the slice-wise characteristics derived from 5 single models or twenty augmented input versions, respectively). Then, the slice-wise absolute differences between characteristics obtained manually and automatically are averaged across all test slices. The mean, standard deviation (SD) and worstcase of the averaged absolute differences are computed across the 25 single models, 5 ensembles or 25 TTA outputs.

Metric	Training set configuration	Single model		Ensembling averaging		TTA averaging	
		Mean ± SD	Worst	Mean ± SD	Worst	Mean ± SD	Worst
Area [pixels]	C(1,10)	9637 ± 1717	13716	9138 ± 200	9401	9807 ± 1854	14167
Area [pixels]	C(2,5)	7985 ± 844	10341	7672 ± 422	8399	7924 ± 987	10671
Area [pixels]	C(2,100)	9562 ± 1554	14470	8825 ± 237	9182	9439 ± 1447	14036
Area [pixels]	C(8,5)	6139 ± 420	7085	5947 ± 158	6121	6058 ± 451	7270
Area [pixels]	C(8,100)	5886 ± 946	7826	5362 ± 151	5635	5604 ± 865	8079
Nb lacunae	C(1,10)	28.9 ± 5.1	44.2	26.5 ± 0.9	27.7	28.8 ± 5.2	42.2
Nb lacunae	C(2,5)	27.6 ± 2.9	34.8	26.3 ± 1.2	28.5	27.2 ± 3.0	33.7
Nb lacunae	C(2,100)	28.2 ± 4.5	44.1	25.8 ± 0.7	26.6	28.3 ± 4.0	41.1
Nb lacunae	C(8,5)	19.5 ± 0.9	21.6	18.4 ± 0.2	18.8	18.8 ± 0.6	20.3
Nb lacunae	C(8,100)	21.9 ± 1.9	27.1	19.8 ± 0.6	20.4	21.6 ± 1.8	27.3
Porosity [%]	C(1,10)	1.23% ± 0.08%	1.5%	1.17% ± 0.01%	1.2%	1.20% ± 0.09%	1.4%
Porosity [%]	C(2,5)	1.10% ± 0.04%	1.2%	1.04% ± 0.01%	1.1%	1.09% ± 0.06%	1.2%
Porosity [%]	C(2,100)	1.20% ± 0.11%	1.5%	1.12% ± 0.06%	1.2%	1.20% ± 0.12%	1.5%
Porosity [%]	C(8,5)	1.12% ± 0.04%	1.2%	1.08% ± 0.02%	1.1%	1.10% ± 0.03%	1.1%
Porosity [%]	C(8,100)	1.34% ± 0.15%	1.7%	1.30% ± 0.08%	1.4%	1.33% ± 0.17%	1.8%

Table 4
Absolute difference between the cartilage characteristics (i.e. area, number of lacunae and porosity) measured for the automatic and manual segmentation approaches. 25 single models have been trained with different initializations. 5 ensembles have been built, each of them considering 5 single models. TTA was applied on every single model, considering twenty augmented versions of the input. Cartilage characteristics are computed slice-wise. Ensembling and TTA consider a fusion at the segmentation level (using pixel-wise majority voting to derive a single segmentation mask and compute slice-wise characteristics from it). Then, the slice-wise absolute differences between characteristics obtained manually and automatically are averaged across all test slices. The mean, standard deviation (SD) and worstcase of the averaged absolute differences are computed across the 25 single models, 5 ensembles or 25 TTA outputs. MV: majority voting.

Metric	Training set configuration	Single model		Ensembling MV		TTA MV	
		Mean ± SD	Worst	Mean ± SD	Worst	Mean ± SD	Worst
Area [pixels]	C(1,10)	9637 ± 1717	13716	9327 ± 213	9652	10156 ± 1914	14562
Area [pixels]	C(2,5)	7985 ± 844	10341	7878 ± 415	8553	8164 ± 1040	11108
Area [pixels]	C(2,100)	9562 ± 1554	14470	8718 ± 226	8951	9579 ± 1238	13856
Area [pixels]	C(8,5)	6139 ± 420	7085	5970 ± 156	6141	6248 ± 473	7493
Area [pixels]	C(8,100)	5886 ± 946	7826	5327 ± 168	5615	5671 ± 926	8222
Nb lacunae	C(1,10)	28.9 ± 5.1	44.2	26.4 ± 0.8	27.7	29.7 ± 5.5	43.7
Nb lacunae	C(2,5)	27.6 ± 2.9	34.8	26.7 ± 1.1	28.6	28.0 ± 3.3	35.2
Nb lacunae	C(2,100)	28.2 ± 4.5	44.1	26.4 ± 1.0	27.8	29.7 ± 3.7	41.7
Nb lacunae	C(8,5)	19.5 ± 0.9	21.6	18.7 ± 0.5	19.1	19.6 ± 0.5	20.7
Nb lacunae	C(8,100)	21.9 ± 1.9	27.1	20.5 ± 0.7	21.3	21.9 ± 1.8	27.4
Porosity [%]	C(1,10)	1.23% ± 0.08%	1.5%	1.16% ± 0.01%	1.2%	1.19% ± 0.06%	1.4%
Porosity [%]	C(2,5)	1.10% ± 0.04%	1.2%	1.08% ± 0.01%	1.1%	1.11% ± 0.05%	1.2%
Porosity [%]	C(2,100)	1.20% ± 0.11%	1.5%	1.17% ± 0.06%	1.3%	1.25% ± 0.12%	1.5%
Porosity [%]	C(8,5)	1.12% ± 0.04%	1.2%	1.12% ± 0.01%	1.1%	1.13% ± 0.03%	1.2%
Porosity [%]	C(8,100)	1.34% ± 0.15%	1.7%	1.33% ± 0.06%	1.4%	1.38% ± 0.16%	1.8%

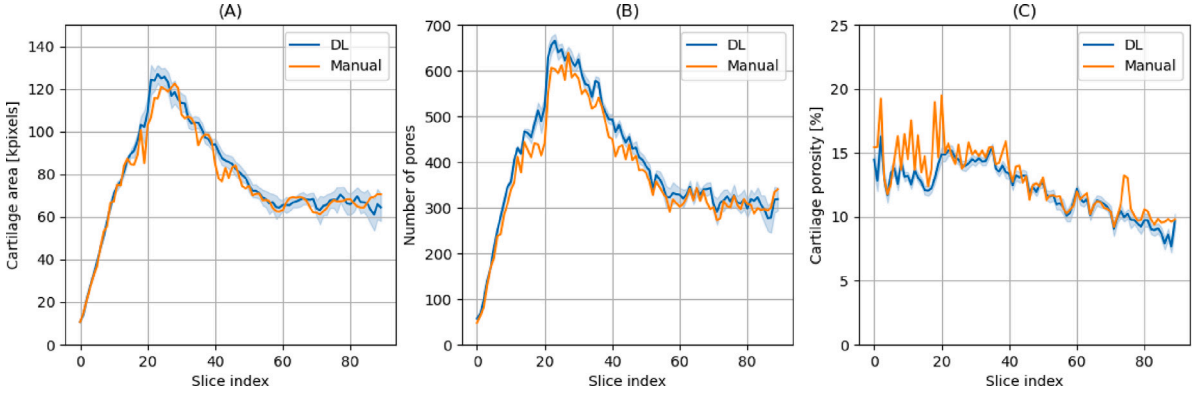

Fig. 3. Cartilage characteristics ((A) area, (B) number of lacunae, or (C) porosity) obtained from the manual and automatic segmentations on a test stack. For the automatic deep learning-based segmentation, twenty models were trained with different initializations. The light blue interval gives the standard deviation of the cartilage characteristics across the twenty models.

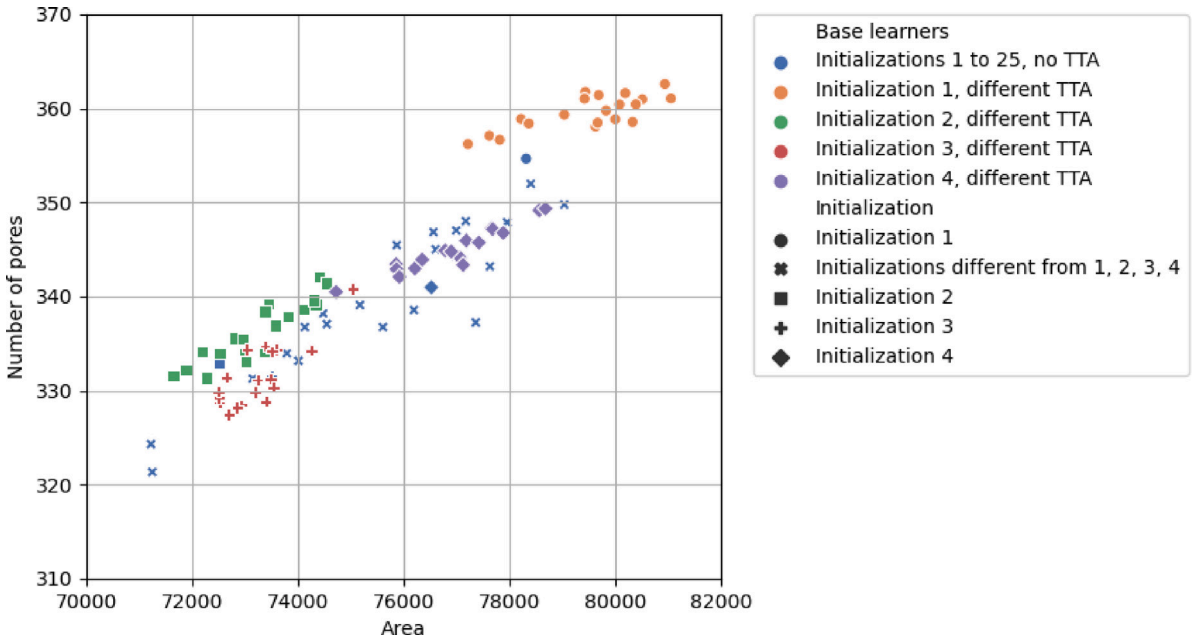

Fig. 4. Slice-wise-determined cartilage area and number of lacunae (or closed pores), as obtained with training configuration $C(8, 100)$. The characteristics obtained with different test-time augmentations are less spread than the predictions obtained with different initializations.

Table 5
Expected calibration error (ECE) for different training set configurations. ECE is computed for a single model, the ensemble, and TTA. ECE decreases when calibration improves.

Training set configuration	Single model ECE %	Ensemble ECE %	TTA ECE %
C(1,10)	7.054	1.243	1.315
C(2,5)	7.059	1.004	1.324
C(2,100)	7.057	0.882	1.120
C(8,5)	7.041	0.829	0.945
C(8,100)	7.027	0.709	0.774

table, the absolute differences computed at slice level were averaged across all the test slices to define a so-called sum of absolute difference (SAD). For each characteristic, the distribution of SADs resulting from distinct models, trained from distinct random initializations, was defined in terms of its mean, standard deviation, and worst case value (see the single model columns of Tables 3 and 4). In addition, the ensembling or TTA strategy was considered to derive a more accurate estimation of the characteristics in each slice before computing their SADs. The combination of the set of cartilage masks resulting from TTA or ensembling consists in either averaging the characteristic evaluation derived from each individual segmentation (Table 3) or computing the characteristic from the mask obtained with majority voting (Table 4). SADs are computed for multiple sets of models, so that the SAD mean, standard deviation, and worst cases can be presented in Tables 3 and 4, to be compared with the single model case.

Tables 3 and 4 show that the mean absolute errors on the area and number of lacunae decrease with the number of stacks in the training set for the single models. The number of stacks benefits the cartilage characterization more significantly than the number of images. This is consistent with the results obtained with the DSC in Table 2. In contrast, porosity estimation does not improve with the number of stacks. We explain this by the fact that porosity is sensitive to false-positive samples (with cartilage being the positive class) but not to false-negative samples. Hence, training configurations with few false-positive samples can yield low porosity estimate errors despite poor segmentation accuracy.

For all training configurations, the standard deviation and mean of the absolute error decrease with ensembling in comparison with single models, whereas TTA does not allow their reduction. This results in a significant improvement in the worst case cartilage characteristic with ensembling compared with single models. We also see that the larger the standard deviation across different initializations, the larger the benefit of ensembling. We explain the accuracy improvement brought by ensembling by the fact that predictions obtained from models trained with different initializations lead to errors that are relatively inconsistent across models. In contrast, TTA is conditioned by the model initialization and the errors are too correlated, in the predictions obtained from different augmented versions of the input, to improve the segmentation accuracy. This explanation is further supported by Fig. 4, which provides a scatter plot of the cartilage area and number of lacunae estimated from predictions obtained with different model initializations and test-time augmentations. We can see that the cartilage characteristics obtained with different initializations are more spread than the predictions obtained with different TTAs, but a single initialization. We also see that the cartilage characteristics obtained with different TTAs remain near the cartilage characteristics obtained with the same initialization but without TTA.

4.2. Model uncertainty

4.2.1. Calibration

The ECE has been computed to determine whether the pixel-wise estimated probabilities, $\hat{p}_i$, were calibrated using a single model, ensembling, or TTA. In the case of a single model, the estimated probability $\hat{p}_i$, was directly the output of the sigmoid, which was part of the model architecture. When ensembling or TTA was used the estimated probability $\hat{p}_i$ was the average of the different predictions. ECE has been computed using pixels of the 290 test slices and is shown in Table 5. Twenty models with different initializations were used for the computation of $\hat{p}_i$ with ensembling and twenty different augmentations were used for the computation of $\hat{p}_i$ with TTA. TTA was applied to the model that was used to obtain the results in the single model case. This model was randomly chosen from the models trained with different initializations. Table 5 shows that both ensembling and TTA improve the calibration of the predicted cartilage probability maps compared with a single model, according to ECE and all training configurations. Also, ensembling provides slightly better-calibrated probability maps than

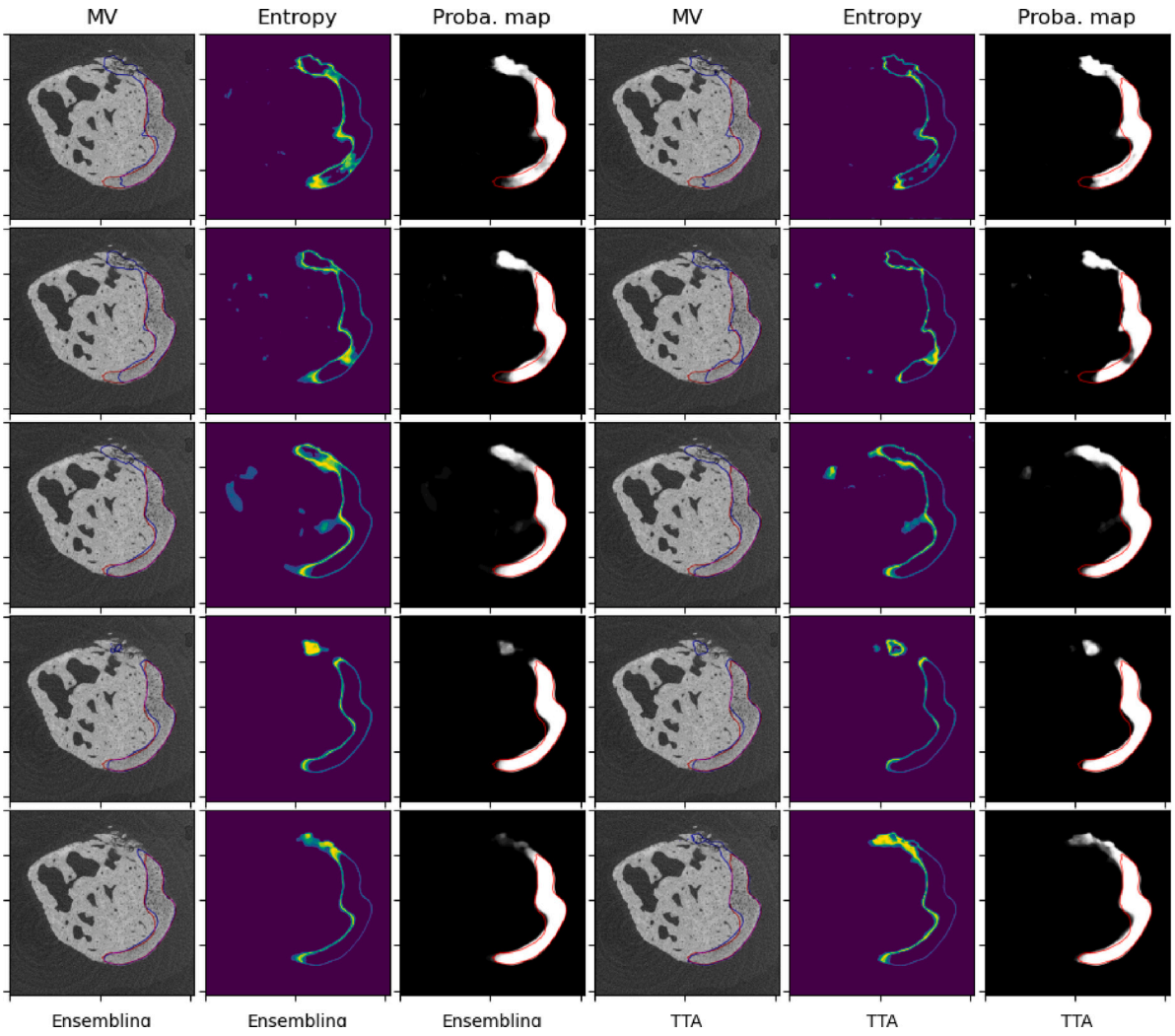

Fig. 5. Qualitative representation of the uncertainty in the model predictions. Pixel-wise entropy and probability maps were computed using twenty different predictions. Red: manual segmentation, blue: automatic segmentation obtained after majority voting (MV) using the twenty predictions.

TTA. However, the difference between ECE obtained with ensembling and TTA is small in relation to the calibration improvement compared with the single model. Hence, TTA can be an interesting alternative to ensembling for calibration, since it produces significant improvements over a single model without requiring training and storing multiple models.

Importantly, calibration and accuracy are orthogonal concerns [22]. Hence, it is desirable to obtain models that are both accurate and calibrated. For such models, an estimated probability close to 0.5 can be explained only by the presence of either ambiguous or out-of-distribution regions in the test set.[3] This is because the high accuracy of the model means that it generalizes correctly to patterns that are largely represented and consistently annotated in the training set. This observation leads us to the conclusion that the calibration improvement observed for most accurate models (trained with configuration $C(8,5)$ or $C(8,100)$) reflects their strong ability to identify ambiguous and out-of-distribution regions, both with ensembling and TTA. This conclusion can be validated qualitatively in Fig. 5. Indeed, for configurations $C(8,5)$ and $C(8,100)$, regions with high pixel-wise entropy are mostly located at the border between cartilage and bone (ambiguous region) or in a region where bone presents pieces that are disconnected from the main bone structure, unlike in the training set (out-of-distribution region). Interestingly, ensembling provides more regions with a high entropy than TTA, especially for training configurations with few stacks (e.g., $C(1,10)$). This is consistent with the fact that ensembling better captures uncertainty linked to the model inaccuracy (due to a small training set) because it can explore different modes through different initializations.

4.2.2. Failure detection

Well-calibrated models are expected to produce uncertain predictions (i.e., estimated probability near 0.5) in regions where cartilage and bone are hard to distinguish. Hence, regions with high prediction uncertainty are expected to be correlated with poor segmentation performances, as measured by the DSC. The correlation between uncertainty and segmentation performance was evaluated slice-wise, to assess whether high uncertainty provides a reliable cue to identify slices in which automatic segmentation needs to be corrected manually by an expert. Therefore, the slice-wise entropy level, s_H, was computed together with the slice-wise majority vote DSC and reported in scatter plots in Fig. 6 for ensembling and TTA. Scatter plots were not computed for the single model since the entropy level is close to zero on every slice due to the model's overconfidence. The Pearson correlation coefficient, r, between the entropy score and DSC is also reported in

[3] Ambiguous regions refer to patterns for which the annotations are inconsistent in the test set because of the difficulty of distinguishing mineralized cartilage from bone even for the annotator. In contrast, out-of-distribution regions correspond to patterns that were not seen during the training phase.

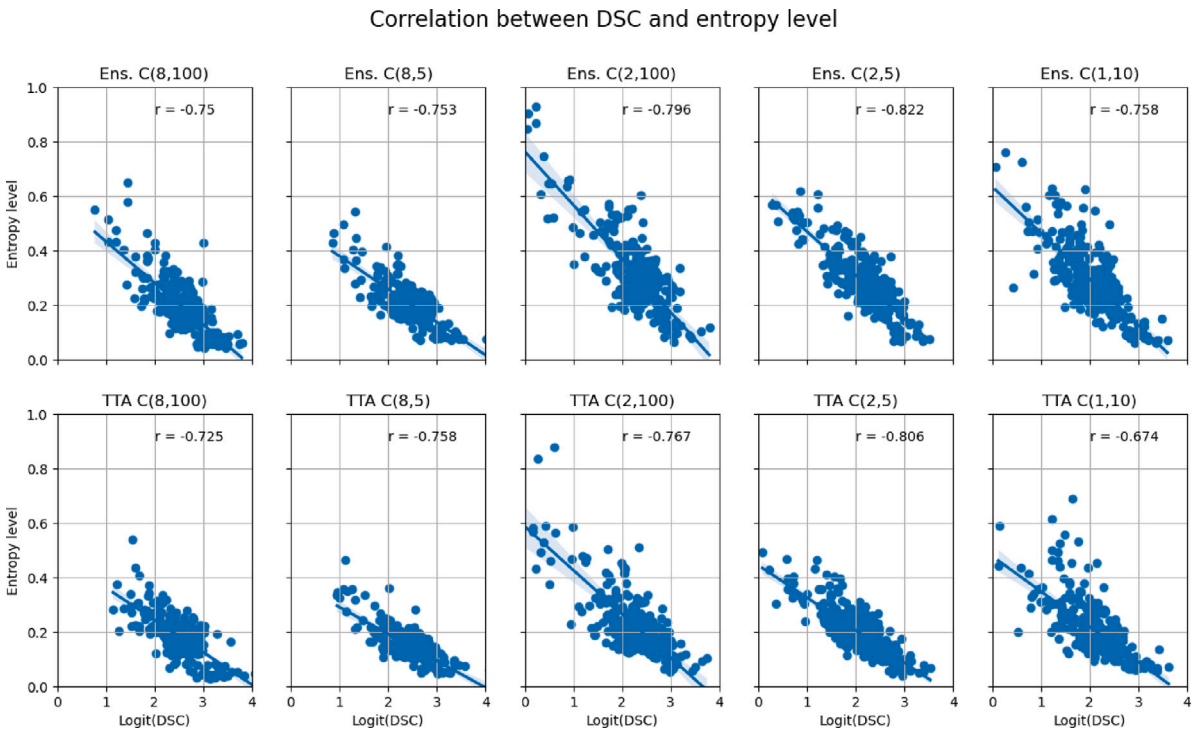

Fig. 6. Slice-wise entropy level as a function of the slice-wise majority vote DSC for all the test slices. For better visualization, Dice values were logit transformed: logit(DSC) = ln(DSC/(1-DSC)) as in [34].

Fig. 6. The Pearson correlation coefficient between random variables A and B is defined by

$$r = \frac{cov(A, B)}{\sigma_A \sigma_B}, \tag{10}$$

where σ_A and σ_B are the standard deviations of A and B, respectively. In Fig. 6, the slice-wise entropy level and slice-wise DSC are shown to be highly correlated ($0.7 < |r|$) for almost all the training configurations and both ensembling and TTA. Only training configuration $C(1, 10)$ used with TTA shows a moderate correlation ($0.5 < |r| < 0.7$). This means that the slice-wise entropy level measured on the probability maps generated with ensembling and TTA can be used as a metric to identify the slices where the segmentation performed poorly.

5. Discussion

Despite the similar grayscale density attenuation of bone and mineralized cartilage, the proposed approach is able to segment mineralized cartilage from bone automatically with a DSC higher than 0.9 when 800 slices are used in the training set. This shows that 2D U-Net is a powerful tool for automatic segmentation of mineralized cartilage versus bone in microCT images. The advantage of the approach is that it is fully automatic and therefore speeds up the segmentation process. We expect this to allow the study of larger numbers of samples and a more representative characterization of mineralized cartilage from a statistical point of view. A limitation of our approach is that we expect poor generalization capabilities to mineralized cartilage segmentation of anatomical sites other than the Achilles tendon–bone interface, since other sites might be significantly different in size and shape. Nevertheless, there is no reason for the proposed approach not to work when training sets that are specific to these new applications become available. Hence, our approach could also be valuable for the segmentation of mineralized cartilage in other bone-to-tendon interfaces, in the osteochondral interface (i.e., the bone–cartilage interface), or in endochondral bone formation.

Interestingly, our work provides valuable recommendations regarding the annotation of samples corresponding to novel use cases. We showed that, for a fixed number of annotations, 2D microCT slices should preferably be selected from distinct 3D microCT volumes, at regular intervals, rather than being grouped in adjacent slices of a same 3D sample. We also showed that model ensembling was able to increase the robustness of the model to unlucky initializations with a small number of labeled training samples, thereby improving the accuracy of segmentation and the assessment of mineralized cartilage features. A limitation of our study is that it focused on the comparison of multiple strategies to increase the reliability of cartilage feature evaluation within a single population only. Future work should investigate whether the performance improvements brought by model ensembling have a significant impact in applicative contexts where the mineralized cartilage characteristics vary across populations depending on a property such as diseases or aging. We also showed that both model ensembling and TTA allow associating confidence ratings to the model's decisions. A straightforward implication is that the prediction reliability estimation helps in identifying samples for which a manual check is worthwhile. In this case, the segmentation approach would be semi-automatic.

Finally, some limitations are common to many deep learning-based approaches. Namely, our approach should be adapted to deal with larger volumes while staying at full resolution because the GPU's memory is limited. This requires a strategy to extract image patches at training. In this case, the inference will also be done patch-based with the same patch size as used during training. To prevent stitching artifacts, adjacent predictions should overlap. Another limitation is that the type of uncertainty (aleatoric versus epistemic) that is captured with TTA remains unclear. This question should be further studied, for instance by considering different augmentations during training and inference. Besides this, the reason why TTA improves model calibration but not segmentation accuracy, in contrast to model ensembling, is not analyzed in detail and requires additional research.

6. Conclusion

Model ensembling and test-time augmentation (TTA) were used to improve the accuracy of mineralized cartilage segmentation versus

bone, assessment of the structural characteristics of the cartilage, and estimation of the model prediction uncertainty for different training configurations. Given a limited amount of annotation resources, we found that slices to annotate should be sampled from distinct 3D volumes, at regular spatial interval (to maximize the distance between annotated slices). Compared with single models, model ensembling was shown to improve DSC and reduce the mean and standard deviation of the absolute error on the cartilage characteristics across different initializations. In contrast, TTA did not improve the prediction accuracy. However, model ensembling and TTA both improved the calibration of the model over single models, according to ECE. Ensembling provides even better-calibrated predictions than TTA, in agreement with the fact that different initializations explore different modes (i.e., learned functions). Calibrated predictions can further be used to predict the slices where the segmentation fails. Indeed, a strong correlation was observed between the prediction inconsistency, captured by the slice-wise averaged entropy of the prediction, and the slice-wise DSC.

CRediT authorship contribution statement

Jean Léger: Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Lisa Leyssens:** Data curation, Writing – review & editing. **Greet Kerckhofs:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Christophe De Vleeschouwer:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Gabrielle Leyden for editing the final revision of this paper and Victor Joos for the discussions about the revision of the paper. We also thank Carla Geeroms for performing the microCT data acquisition.

Funding

Jean Léger is a Research Fellow with Belgium's National Scientific Research Foundation, Belgium and Christophe De Vleeschouwer is Research Director with Belgium's National Scientific Research Foundation. Lisa Leyssens is funded by an FSR project - Projets Fonds de Recherche Spécial (FSR) Jeunes Académiques - Fédération Wallonie-Bruxelles, Belgium and the UCLouvain, Belgium (assistant position). This research project was funded by a research project of the Flanders Research Foundation, Belgium (FWO; Grant No. G088218N). The high-resolution micro-CT images were generated at the X-ray computed tomography facilities of the Department of Development and Regeneration of KU Leuven, financed by the Hercules Foundation, Belgium (project AKUL 13/47). The project is also supported by the Action de Recherche Concertée (ARC 19/24-097 – Bio-Blueprints) – Fédération Wallonie-Bruxelles, Belgium.

References

[1] G. Kerckhofs, J. Sainz, M. Maréchal, M. Wevers, T. Van de Putte, L. Geris, J. Schrooten, Contrast-enhanced nanofocus X-ray computed tomography allows virtual three-dimensional histopathology and morphometric analysis of osteoarthritis in small animal models, Cartilage 5 (1) (2014) 55–65.

[2] G. Kerckhofs, J. Sainz, M. Wevers, T. Van de Putte, J. Schrooten, Contrast-enhanced nanofocus computed tomography images the cartilage subtissue architecture in three dimensions, Eur. Cells Mater. 25 (2013) 179–189.

[3] H.H. Lu, S. Thomopoulos, Functional attachment of soft tissues to bone: development, healing, and tissue engineering, Annu. Rev. Biomed. Eng. 15 (2013) 201–226.

[4] S.J. Rytky, L. Huang, P. Tanska, A. Tiulpin, E. Panfilov, W. Herzog, R.K. Korhonen, S. Saarakkala, M.A. Finnilä, Automated analysis of rabbit knee calcified cartilage morphology using micro-computed tomography and deep learning, J. Anat. (2021).

[5] R. Fu, C. Liu, Y. Yan, Q. Li, R.-L. Huang, Bone defect reconstruction via endochondral ossification: a developmental engineering strategy, J. Tissue Eng. 12 (2021) 20417314211004211.

[6] L. Si, K. Xuan, J. Zhong, J. Huo, Y. Xing, J. Geng, Y. Hu, H. Zhang, Q. Wang, W. Yao, Knee cartilage thickness differs alongside ages: a 3-T magnetic resonance research upon 2,481 subjects via deep learning, Front. Med. 7 (2021) 600049.

[7] A. Cheng, Z. Schwartz, A. Kahn, X. Li, Z. Shao, M. Sun, Y. Ao, B.D. Boyan, H. Chen, Advances in porous scaffold design for bone and cartilage tissue engineering and regeneration, Tissue Eng. B 25 (1) (2019) 14–29.

[8] T.R. Oegema, R.J. Carpenter, F. Hofmeister, R.C. Thompson, The interaction of the zone of calcified cartilage and subchondral bone in osteoarthritis, Microsc. Res. Tech. 37 (4) (1997) 324–332.

[9] H.-S. Gan, M.H. Ramlee, A.A. Wahab, Y.-S. Lee, A. Shimizu, From classical to deep learning: review on cartilage and bone segmentation techniques in knee osteoarthritis research, Artif. Intell. Rev. (2020) 1–50.

[10] J. Lian, Z. Yang, J. Liu, W. Sun, L. Zheng, X. Du, Z. Yi, B. Shi, Y. Ma, An overview of image segmentation based on pulse-coupled neural network, Arch. Comput. Methods Eng. 28 (2) (2021) 387–403.

[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[12] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[13] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211.

[14] H. Zheng, S.M.M. Perrine, M.K. Pitirri, K. Kawasaki, C. Wang, J.T. Richtsmeier, D.Z. Chen, Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 802–812.

[15] J. Léger, L. Leyssens, C. De Vleeschouwer, G. Kerckhofs, Deep learning-based segmentation of mineralized cartilage and bone in high-resolution micro-CT images, in: International Symposium on Computer Methods in Biomechanics and Biomedical Engineering. Springer, 2019, pp. 158–170.

[16] I. Papantoniou, M. Sonnaert, L. Geris, F.P. Luyten, J. Schrooten, G. Kerckhofs, Three-dimensional characterization of tissue-engineered constructs by contrast-enhanced nanofocus computed tomography, Tissue Eng. C 20 (3) (2014) 177–187.

[17] L. Leyssens, C. Pestiaux, G. Kerckhofs, A review of ex vivo X-ray microfocus computed tomography-based characterization of the cardiovascular system, Int. J. Mol. Sci. 22 (6) (2021) 3263.

[18] J. Léger, E. Brion, P. Desbordes, C. De Vleeschouwer, J.A. Lee, B. Macq, Cross-domain data augmentation for deep-learning-based male pelvic organ segmentation in cone beam CT, Appl. Sci. 10 (3) (2020) 1154.

[19] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J.N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation, Med. Image Anal. 63 (2020) 101693.

[20] S. Fort, H. Hu, B. Lakshminarayanan, Deep ensembles: A loss landscape perspective, 2019, arXiv preprint arXiv:1912.02757.

[21] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On calibration of modern neural networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1321–1330.

[22] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, 2016, arXiv preprint arXiv:1612.01474.

[23] D.J. MacKay, Bayesian Methods for Adaptive Models (Ph.D. thesis), California Institute of Technology, 1992.

[24] R.M. Neal, Bayesian Learning for Neural Networks, Vol. 118, Springer Science & Business Media, 2012.

[25] A. Graves, Practical variational inference for neural networks, in: Advances in Neural Information Processing Systems, Citeseer, 2011, pp. 2348–2356.

[26] C. Louizos, M. Welling, Structured and efficient variational deep learning with matrix gaussian posteriors, in: International Conference on Machine Learning, PMLR, 2016, pp. 1708–1716.

[27] M. Teye, H. Azizpour, K. Smith, Bayesian uncertainty estimation for batch normalized deep networks, in: International Conference on Machine Learning, PMLR, 2018, pp. 4907–4916.

[28] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.

[29] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision? 2017, arXiv preprint arXiv:1703.04977.

[30] M.S. Ayhan, P. Berens, Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks, 2018.

[31] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34–45.
[32] D. Karimi, Q. Zeng, P. Mathur, A. Avinash, S. Mahdavi, I. Spadinger, P. Abolmaesumi, S.E. Salcudean, Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images, Med. Image Anal. 57 (2019) 186–196.
[33] A.-J. Rousseau, T. Becker, J. Bertels, M.B. Blaschko, D. Valkenborg, Post training uncertainty calibration of deep networks for medical image segmentation, 2020, arXiv preprint arXiv:2010.14290.
[34] A. Mehrtash, W.M. Wells, C.M. Tempany, P. Abolmaesumi, T. Kapur, Confidence calibration and predictive uncertainty estimation for deep medical image segmentation, IEEE Trans. Med. Imaging 39 (12) (2020) 3868–3878.
[35] M. Ng, F. Guo, L. Biswas, S.E. Petersen, S.K. Piechnik, S. Neubauer, G. Wright, Estimating uncertainty in neural networks for cardiac MRI segmentation: A benchmark study, 2020, arXiv preprint arXiv:2012.15772.
[36] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, M. Nielsen, Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 246–253.
[37] F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, R. Kijowski, Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging, Magn. Reson. Med. 79 (4) (2018) 2379–2391.
[38] B. Norman, V. Pedoia, S. Majumdar, Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry, Radiology 288 (1) (2018) 177–185.
[39] Z. Zhou, G. Zhao, R. Kijowski, F. Liu, Deep convolutional neural network for segmentation of knee joint anatomy, Magn. Reson. Med. 80 (6) (2018) 2759–2770.
[40] F. Ambellan, A. Tack, M. Ehlke, S. Zachow, Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative, Med. Image Anal. 52 (2019) 109–118.
[41] A. Raj, S. Vishwanathan, B. Ajani, K. Krishnan, H. Agarwal, Automatic knee cartilage segmentation using fully volumetric convolutional neural networks for evaluation of osteoarthritis, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 851–854.
[42] E. Peake, R. Chevasson, S. Pszczolkowski, D.P. Auer, C. Arthofer, Ensemble learning for robust knee cartilage segmentation: data from the osteoarthritis initiative, BioRxiv (2020).
[43] M. Antico, F. Sasazawa, Y. Takeda, A.T. Jaiprakash, M.-L. Wille, A. Pandey, R. Crawford, G. Carneiro, D. Fontanarosa, Bayesian CNN for segmentation uncertainty inference on 4D ultrasound images of the femoral cartilage for guidance in robotic knee arthroscopy, IEEE Access (2020).
[44] P.R. Tabrizi, R.A. Zoroofi, F. Yokota, S. Tamura, T. Nishii, Y. Sato, Acetabular cartilage segmentation in CT arthrography based on a bone-normalized probabilistic atlas, Int. J. Comput. Assist. Radiol. Surg. 10 (4) (2015) 433–446.
[45] K.A. Myller, J.T. Honkanen, J.S. Jurvelin, S. Saarakkala, J. Töyräs, S.P. Väänänen, Method for segmentation of knee articular cartilages based on contrast-enhanced CT images, Ann. Biomed. Eng. 46 (11) (2018) 1756–1767.
[46] J. Apostolakos, T.J. Durant, C.R. Dwyer, R.P. Russell, J.H. Weinreb, F. Alaee, K. Beitzel, M.B. McCarthy, M.P. Cote, A.D. Mazzocca, The enthesis: a review of the tendon-to-bone insertion, Muscles Ligaments Tendons J. 4 (3) (2014) 333.
[47] A. Karim, A.K. Amin, A.C. Hall, The clustering and morphology of chondrocytes in normal and mildly degenerate human femoral head cartilage studied by confocal laser scanning microscopy, J. Anat. 232 (4) (2018) 686–698.
[48] O. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 424–432.
[49] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Icml, 2010.
[50] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[51] M.P. Naeini, G. Cooper, M. Hauskrecht, Obtaining well calibrated probabilities using bayesian binning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 29, 2015.
[52] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 625–632.
[53] A. Jungo, M. Reyes, Assessing reliability and challenges of uncertainty estimations for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 48–56.