IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 23, 2020, accepted December 8, 2020, date of publication December 14, 2020,
date of current version December 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3044355

Bayesian CNN for Segmentation Uncertainty Inference on 4D Ultrasound Images of the Femoral Cartilage for Guidance in Robotic Knee Arthroscopy

MARIA ANTICO[1,2], FUMIO SASAZAWA[3,4,5], YU TAKEDA[6], ANJALI TUMKUR JAIPRAKASH[2,7], MARIE-LUISE WILLE[1,2], (Member, IEEE), AJAY K. PANDEY[2,8], ROSS CRAWFORD[1,2], GUSTAVO CARNEIRO[9], AND DAVIDE FONTANAROSA[2,7]

[1]Science and Engineering Faculty, School of Mechanical Medical and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia
[2]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD 4000, Australia
[3]Department of Orthopaedic Surgery, Faculty of Medicine, Hokkaido University, Sapporo 060-0808, Japan
[4]Graduate School of Medicine, Hokkaido University, Sapporo 060-0808, Japan
[5]Department of Orthopaedic Surgery, Hakodate Central General Hospital, Sapporo 040-8585, Japan
[6]Department of Orthopedic Surgery, Hyogo College of Medicine, Nishinomiya 663-8501, Japan
[7]School of Clinical Sciences, Queensland University of Technology, Brisbane, QLD 4000, Australia
[8]Science and Engineering Faculty, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia
[9]School of Computer Science, Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA 5005, Australia

Corresponding author: Davide Fontanarosa (d3.fontanarosa@qut.edu.au)

This work was supported by the Australia-India Strategic Research Fund (Intelligent Robotic Imaging System for Keyhole Surgeries) under Grant AISRF53820.

ABSTRACT Ultrasound (US) imaging is a complex imaging modality, where the tissues are typically characterised by an inhomogeneous image intensity and by a variable image definition at the boundaries that depends on the direction of the incident sound wave. For this reason, conventional image segmentation approaches where the regions of interest are represented by exact masks are inherently inefficient for US images. To solve this issue, we present the first application of a Bayesian convolutional neural network (CNN) based on Monte Carlo dropout on US imaging. This approach is particularly relevant for quantitative applications since differently from traditional CNNs, it enables to infer for each image pixel not only the probability of being part of the target but also the algorithm confidence (i.e. uncertainty) in assigning that probability. In this work, this technique has been applied on US images of the femoral cartilage in the framework of a new application, where high-refresh-rate volumetric US is used for guidance in minimally invasive robotic surgery for the knee. Two options were explored, where the Bayesian CNN was trained with the femoral cartilage contoured either on US, or on magnetic resonance imaging (MRI) and then projected onto the corresponding US volume. To evaluate the segmentation performance, we propose a novel approach where a probabilistic ground-truth annotation was generated combining the femoral cartilage contours from registered US and MRI volumes. Both cases produced a significantly better segmentation performance when compared against traditional CNNs, achieving a dice score coefficient increase of about 6% and 8%, respectively.

INDEX TERMS 4D ultrasound, Bayesian CNN, deep learning, MRI-US registration, robotic knee arthroscopy, ultrasound guided minimally invasive surgery, ultrasound guided arthroscopy, ultrasound guidance, uncertainty.

I. INTRODUCTION

Ultrasound (US) is broadly used to scan many body regions (e.g. abdomen, musculoskeletal system) due to its capability to visualise both bony surfaces and soft tissues. Moreover, it has several advantages over other imaging modalities such as cost-effectiveness, portability, non-invasiveness and volumetric ''real-time'' (high-refresh-rate) capability, making this imaging modality appealing for many applications. Despite these facts, US imaging is currently not exploited at its full

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

potential due to challenging aspects in image interpretation. While other imaging modalities, such as Magnetic resonance imaging (MRI) and CT, capture the anatomical information scanning the region of interest (ROI) from different directions, US is a ''mono-directional'' modality: it utilises one line of view to image the tissues. Due to the physics of US imaging, whenever the line of view is not perpendicular to the surface of the tissue, the reflection from the tissue surface is partially deflected from the US probe resulting in a weaker signal. As a consequence, the tissue would be characterised by an inhomogeneous image intensity and by not well-defined boundaries on the image generated. For this reason, tissue interfaces on US images typically cannot be represented by a sharp line, and thus conventional image segmentation approaches where the ROIs are represented by exact masks are inherently inefficient for US images. This is critical especially for quantitative applications such as surgical guidance, where it is essential to generate an accurate tissue representation, including all the regions potentially belonging to the ROI.

In this work, we propose a solution to this issue in the framework of a novel application currently investigated by our group, where high-refresh-rate volumetric US (referred to as to 4D US or 3DUS+time) is used for guidance in minimally invasive robotic surgery for the knee [1], [2]. The femoral cartilage is the structure most commonly at risk during this procedure and as such it requires a particularly accurate identification [3], [4]. Deep learning (DL) algorithms and convolutional neural networks (CNNs) for automatic image analysis hold potential to deal with the complexity of US imaging [5]–[7]. In our previous works, CNNs were implemented to detect, segment and track the femoral cartilage with clinical accuracy [8]–[11]. As for all traditional deep learning (DL) models, these CNNs associate a prediction to each image pixel with a deterministic approach. These predictions are not fully representative of all the image information provided, in particular for those image regions that are not sufficiently defined to be confidently classified as either part of the target or the background.

Herein, we aim at enhancing this tissue representation by training a Bayesian CNN based on Monte Carlo dropout to predict for each image pixel not only the probability of being part of the target (the femoral cartilage) but also the algorithm confidence (i.e. uncertainty) in assigning that probability. This approach was first introduced by [12] and it gained popularity especially in the computer vision field as it can be used for both detection [13], [14] and segmentation tasks [15] on any existing DL algorithm. Several studies reported the use of this technique in medical imaging analysis, mostly for disease segmentation and detection but, to the best of our knowledge, it has never been applied on US images. A number of research groups [16]–[19] used this approach on MRI volumes of the brain, while Leibig *et al.* [20] and Ozdemir *et al.* [21] on fundus images and CTs of the lung, respectively. These works showed that uncertainty estimation can be useful for automatic disease detection to support the clinicians and enable a faster and more reliable clinical workflow. The main quantitative findings reported showed that uncertainty areas correlate with incorrect predictions. More specifically, this performance was assessed by plotting Receiver Operating Characteristic (ROC) curves, where for each curve the pixels corresponding to an uncertainty level above a certain threshold were excluded from the evaluation and proved that, when the uncertain pixels were discarded, the algorithm performance was higher.

This type of assessment has two main limitations, though:

1. The evaluation showing the improved accuracy did not include all the pixels present in the images.
2. The probabilistic predictions provided by the Bayesian CNN are compared to deterministic ground-truths provided by a human annotator, assuming thus that the expert was 100% confident about each pixel contoured while generating the ground-truth, or that multiple annotators would generate exactly the same annotation. This is obviously an unrealistic assumption, even with imaging modalities offering on average superior imaging quality, such as MRI of the brain [22]. For US this assumption is particularly wrong. In fact, typically intra and inter-operator dice score coefficients can be as low as 65%-70% for clinically acceptable segmentations [8].

In this article, we propose a novel approach where a probabilistic ground-truth annotation to evaluate the CNN performance is generated combining the femoral cartilage contours from registered US and MRI volumes. MRI imaging is considered the ''gold standard'' for femoral cartilage diagnosis and provides a comprehensive and well defined (high contrast/resolution) representation of this anatomical structure [23]. We used this information to provide objective evidence of the cartilage presence, overcoming the inter and intra-observer variability for those intrinsically ambiguous areas where the cartilage cannot be confidently detected using US imaging only. Furthermore, we also explored the option of training the Bayesian CNN to segment the femoral cartilage from US volumes using as labels the MRI-based segmentations of the femoral cartilage projected onto the US volumes. This approach was able to recognise a larger range of pixels on the US images belonging to the cartilage, but of course requires that US and MRI datasets are registered.

II. MATERIALS AND METHODS

A. DATA ACQUISITION

The Queensland University of Technology Ethics Committee granted the approval for the US and MRI data acquisitions (No. 1700001110) described in Sections II.A.1 and II.A.2. Informed consent was obtained from all volunteers prior to data collection. A summary of the dataset utilised in this study is reported in Table 1.

1) 4D US SEQUENCE ACQUISITION

Seven volunteers' knees were imaged using a Philips VL13-5 US probe and a Philips EPIQ7 US system (Philips Medical Systems, Andover, MA, United States). A detailed

TABLE 1. 4D US sequences and MRI volumes. Volunteers' information is reported in Columns 1-6. The total number of 4D sequences acquired and whether an MRI was collected (denoted by the checkmark) are reported for each volunteer in Columns 7 and 8, respectively.

Volunteer ID	Sex	Age	Weight [kg] Height [cm]	Leg	Femoral cartilage pathologies	Number of 4D US sequences	MRI
1	M	34	60 171	L, R	-	6	-
2	F	34	64 170	L,R	-	6	✓
3	M	31	71 185	L,R	-	6	-
4	M	44	80 183	L,R	Partial thickness degeneration in both legs	6	-
5	M	34	78 180	L,R	-	6	✓
6	F	20	43 153	L,R	-	5	-
7	F	26	58 170	L	-	3	✓

description of the US scanning protocol and of the US system settings is reported in [8].

In brief, during the US scans the probe was positioned on the patellar tendon and 4D US sequences were dynamically acquired:

- during leg extension from 30 to 0 degrees knee flexion;
- keeping the knee fixed at either 0 or 30 degrees flexion while the probe translated along the caudal direction from the patella tip up to the point where the femoral cartilage was not visible anymore along the patient's sagittal plane.

These three scanning options covered all possible surgical scenarios of knee arthroscopy with the knee in the 0-30 degrees flexion range and were proven to be compatible with this surgical procedure [1]. The 0 degree knee flexion angle was defined as the ''neutral'' leg position or extended leg; while the 30 degree knee flexion was obtained creating a 15 degree angle between the neutral leg position and both the femur and the tibia. A customised leg cushion was utilised to support the leg at the 30 degree knee angle. To avoid possible acoustic coupling discontinuities between the US probe and the knee surface, the volunteers' knees were scanned while submerged in water.

During the 4D US sequences acquisition, 3D US volumes (with a size of approximately $(4 \times 4 \times 3)$ cm^3) were collected with a 1 Hz full volume refresh rate. In total, 38 4D US sequences were acquired including 164 3D US volumes.

2) MRI ACQUISITION

Three of the seven volunteers involved in this study were also imaged using a 3T MRI system (Siemens Magnetom 3T Prisma, Erlangen, Germany) with 3D SPACE sequences in PD-weighting. For each volunteer, an MRI scan was acquired for either the left or the right knee, with the volunteer in the supine position and the extended leg positioned in dedicated knee coils. The voxel spacing in each MRI was isotropic and was either 0.5 or 0.7 mm.

B. GROUND-TRUTH LABELS GENERATION

1) US-BASED GROUND-TRUTH

The femoral cartilage was outlined by an experienced orthopaedic surgeon (F.S.) on all the 3D US volumes of the 4D US sequences acquired using a customised graphical user interface (GUI) created in Mevislab (MeVis Medical Solutions AG, Germany). The contours were drawn on the sagittal slices of the US volumes since this was the highest resolution plane. The resulting segmentations will be referred to as ''US-based ground-truth'' throughout this article.

2) MRI-BASED PSEUDO-GROUND-TRUTH

a: MRI SEGMENTATION

The femoral cartilage was also outlined by an experienced orthopaedic surgeon (Y.T.) on the three acquired MRI volumes. The contours were outlined on the sagittal plane of the MRI utilising the same GUI as in Section II.B.1.

b: REGISTRATION BETWEEN MRI SEGMENTATIONS AND US VOLUMES

The MRI volume of each of the three volunteers was manually registered to the US volumes in the 4D US sequences of the corresponding knee, resulting in a total of 23 MRI-US volume pairs matched. The registration selected was a rigid roto-translation and aimed at overlaying the femoral cartilage in the two modalities. Since this anatomical structure is rigid (it does not deform for different knee flexion angles), it was possible to find a match between the femoral cartilage in the two modalities even for those cases where the US and the MRI volumes were acquired at different knee flexion angles.

The registration procedure was performed by Y.T. using Imfusion (ImFusion, München, Germany). The software allowed the user to rotate/translate the US and MRI volumes in a common reference coordinate system, selecting the translation (in mm) and rotation values (in degrees) of a transformation matrix to be applied to each volume along the three orthogonal planes. The MRI and US volumes were simultaneously visualised (e.g. using colour or checkerboard blending) along the three orthogonal planes and in the 3D rendering mode while the expert modified the transformation matrices.

Initially, the MRI was rotated within the common reference coordinate system such that the three planes in the view panel would match the three anatomical planes. The convention used was to set the coronal plane to intersect the medial and the lateral femur epicondyle; the axial plane aligned with the lateral and medial parts of the trochlear groove and the sagittal plane orthogonal to the other two anatomical planes. The transformation matrix associated with the US volume was

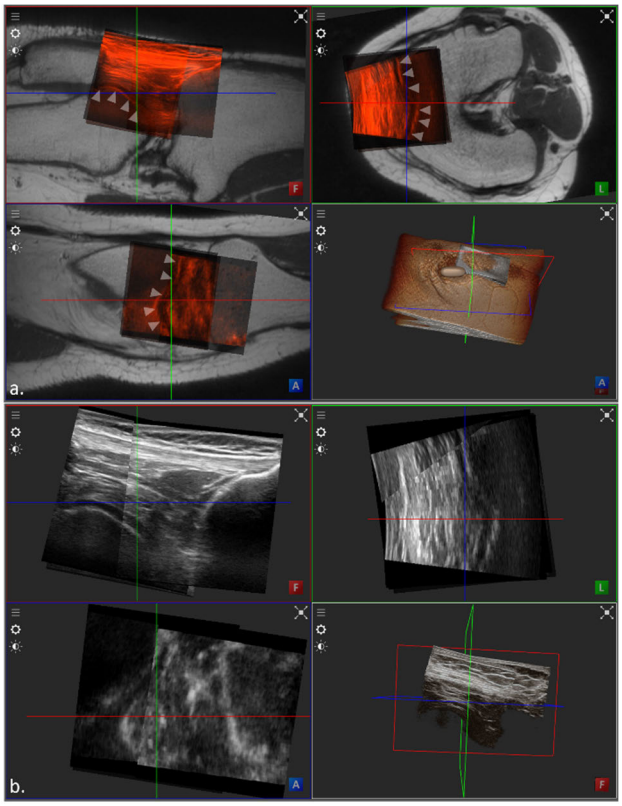

FIGURE 1. MRI-US registration example. a) An MRI volume and multiple 3D US volumes of a 4D sequence registered based on the femoral cartilage (white triangles along the three anatomical planes) overlap in the two modalities. b) The overlap between anatomical areas captured by different US volumes visualised using alpha blending.

then modified such that the femoral cartilage would match between the two modalities along the three anatomical planes (Figure 1a). Thus, the resulting registration purely relied on the image information in the two modalities.

The 3D US volumes of a same 4D US sequence were imported in the software with the corresponding transformations and the overlap between the same anatomical areas captured by different volumes was assessed and improved, in case needed, by further adjusting the US transformations (Figure 1b). Once the MRI-US registration was finalised, for each MRI-US pair registered the US volume was set to its original position, and the coordinates of the registered MRI were modified accordingly.

Finally, for each MRI-US pair, the MRI segmentations were imported in Imfusion and registered to the corresponding US volume using the MRI transformation.

c: MRI-BASED SEGMENTATION POST-PROCESSING

The MRI segmentations were then resampled to the same voxel spacing as the registered US volume through a linear interpolation and smoothed using a kernel radius of 10. Since the MRI segmentation volume was significantly larger than the US volume, the MRI segmentation voxels located outside the US volume field of view were cropped. These post-processing steps were automatically performed through

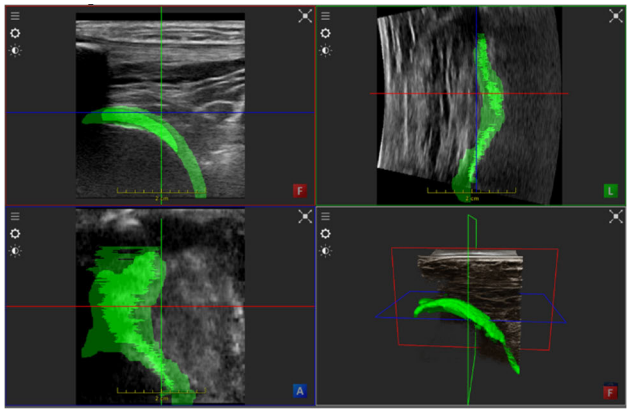

FIGURE 2. Example of MRI segmentation (green alpha-blended) registered to and resampled based on one of the corresponding US volumes. The US-based ground-truth is shown in bright green.

the ''image resampling '' and ''smoothing'' functions embedded in Imfusion. Figure 2 shows an example of MRI segmentation (green alpha-blended) registered to one of the corresponding US volumes and resampled according to the US volume dimensions.

Finally, all the MRI segmented voxels corresponding to the US volume voxels where no US signal was present had to be discarded and considered as part of the background (i.e. segmentation pixel intensity set to '0'). This effect was present for the femoral cartilage aspects shadowed by the patella, as bony surfaces almost completely reflect the sound waves, and for the bottom part of the US volume where padding with black pixels was present. The bone shadowing effect occurred only at the extreme cranial part of the US volumes, where the cartilage was typically perpendicular to the incident sound waves and it was consequently well defined on the US volume. For this reason, we assumed that the most cranial voxels of the femoral cartilage where the US signal was present would be included in the US-based ground-truth. Based on this assumption, for each sagittal slice of the MRI segmentation, the most cranial pixel of the corresponding US-based ground-truth slice was selected to define the margin after which the MRI segmentation pixels would be retained. To solve for the padding problem instead, all the MRI segmentation pixels corresponding to US volume pixels with intensity '0' were considered as background. This was possible since the US volumes contained almost no pixels with intensity value exactly equal to zero, besides the padding regions. Possible holes generated (e.g. 1-2 pixels holes) in the MRI segmentation were then filled using the function 'imfill' in Matlab (Version 9.3.0 (R2018b), The Mathworks Inc. Natick, MA, United States). This final MRI segmentation will be referred to as ''MRI-based pseudo-ground-truth'' throughout this article. The term ''pseudo'' was used since this segmentation was not directly outlined on the US volume based on the US information, but it was the result of projecting the femoral cartilage segmentations from the MRI to the US volume. It should be noted that the segmentations of one MRI were matched through the registration process to all

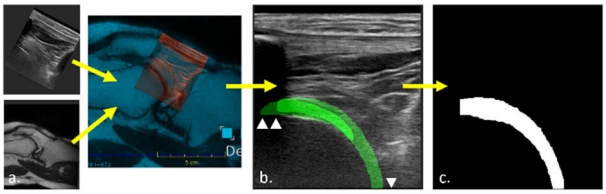

FIGURE 3. Summary of the procedure to create the MRI-based pseudo-ground-truth: a) MRI-US volume pair registered based on the femoral cartilage information; b) MRI segmentation registered to the US volume through the MRI transformation and resampled based on the US voxel/volume dimensions. White triangles highlight the areas included in the MRI segmentation where no signal was present; c) Post-processing of the MRI segmentation to exclude segmentation areas where no US signal is present.

US volumes in the 4D US sequences of the corresponding knee. Figure 3 shows an illustrative summary of the steps followed to create the MRI-based pseudo-ground-truth.

C. 4D US SEQUENCES AND GROUND-TRUTH LABELS POST-PROCESSING

2D US images were obtained by slicing the 3D US volumes of the 4D sequences (Section II.A.1) along the sagittal axis. The resulting 2D US images and the MRI-based pseudo-ground-truths (Section II.B.2), for which the femoral cartilage was not delineated during the US-based ground-truths generation, were discarded. This resulted in a total of 16973 US images, with corresponding US-based ground-truths, of which 3067 had also the corresponding MRI-based pseudo-ground-truth. The 2D US images and all the ground-truth labels were then resized to a pixel size height of 0.09 mm and width of 0.14 mm (the largest pixel dimensions within the dataset). Black pixel padding was applied such that all the images would match the largest image in the dataset (510 pixels x 272 pixels). The size of the images was finally down-sampled to 304 pixels x 160 pixels for faster computation.

D. BAYESIAN CNN FOR SEMANTIC SEGMENTATION

1) THEORY: BAYESIAN CNN WITH MONTE CARLO DROPOUT

Bayesian inference is the most common method utilised to estimate model uncertainty associated with a prediction. Theoretically, this can be achieved by finding the probability distribution (or the posterior distribution) over the weights (W) of the CNN model, given the training dataset of images (X) and corresponding ground-truth labels (Y) (Eq. 1).

$$P(W|X, Y) \tag{1}$$

However, due to the large number of parameters and non-linearities in the model, the posterior distribution is not tractable and thus an approximated distribution $q(W)$ needs to be obtained. In this article, this approximation of distribution was achieved by using Monte Carlo (MC) dropout as proposed by some previous works in literature (e.g [21], [22]). Dropout is a technique typically used during model training to avoid over-fitting [25]. It consists in randomly "dropping out" or switching off a part of the CNN weights based on a predefined probability. The approximated distribution $q(W)$ can be obtained by training a CNN with MC dropout, as proved by Gal and Ghahramani [12]:

- applying dropout after each convolutional layer is equivalent to placing a Bernoulli distribution over each weight of the CNN;
- minimising the commonly used cross-entropy loss using a standard optimisation method is comparable to minimising the Kullback-Leibler (KL) divergence between the approximated and the actual posterior (Eq. 2).

$$KL(q(W)\,||P(W|X, Y)) \tag{2}$$

Once the CNN was trained, samples can be drawn from the posterior distribution $q(W)$. This procedure is done by feeding the images to be segmented to the CNN with MC dropout multiple times (thus each time randomly disabling part of the model weights). As a result, a given image X would be passed n times into slightly different models, from which n predictions Y are obtained. These predictions are then superimposed and the pixel-wise mean prediction and variance computed. The mean prediction $y_{i,mean}$ associated with each pixel i corresponds to the probability of that pixel belonging to the cartilage; and the variance $y_{i,var}$ describes how confident the algorithm was in assigning the corresponding probability (Eqs. 3-4).

$$y_{i,mean} = \frac{1}{n}\sum_{m=1}^{n} y_{i,m} \tag{3}$$

$$y_{i,var} = \frac{1}{n}\sum_{m=1}^{n} (y_{i,m} - y_{i,mean})^2 \tag{4}$$

2) EXPERIMENTS

a: CNN TRAINING WITH US-BASED GROUND-TRUTH

The first type of CNN model was trained to perform semantic segmentation of the femoral cartilage using the US images and the corresponding US-based ground-truths (Sections II.B.1 and II.C). The CNN architecture used was a UNet [26]. The same structure and hyperparameters as in our previous paper were utilised [8], as the same type of images and anatomical region are targeted (Table 2). During training, dropout was applied after each encoder and decoder units, with the probability of discarding network weights set either to 10% or 50%. Differently from other works using MC dropout, we explored the alternative of performing the training with the Dice loss, as it enhanced the CNN performance compared to the commonly used cross-entropy loss.

TABLE 2. UNet Hyperparameters for training with US-based ground-truth.

Hyperparameters	Value
Learning rate	10^{-3}
Weight decay	10^{-5}
Momentum	0.95
Batch size	8
Epochs	31

Two identical models were trained, each time leaving out from the training set the volunteers for testing. The latter consisted of the three volunteers with both US-based ground-truths and the MRI-based pseudo-ground-truths provided (Sections II.B.1 and II.B.2). The two models were trained on 14305 and 15620 US labelled images, respectively; and tested on 2668 and 2324 US images, respectively, obtained after post-processing as described in Section II.C.

b: CNN TRAINING WITH MRI-BASED PSEUDO-GROUND-TRUTH

A new CNN model with the same structure as in Section II.D.2 was trained with the MRI-based pseudo-ground-truths from two volunteers (2324 US labelled images). Due to the limited dataset available, transfer learning from one of the two models described in the previous section was used to initialise the model weights. It should be noted that to avoid any possible bias the model selected for transfer learning was not generated with the data from the subject here utilized at test-time. The hyperparameters selected for training listed in Table 2 remained unchanged apart from the learning rate and the number of training epochs, that were set to 10^{-5} and 10, respectively. The same dropout configuration and probabilities of discarding network weights as in Section II.D.2a were utilised. The model was tested on 743 images of 1 volunteer (that was not included in the training set).

c: BAYESIAN CNNs AND BASELINE TRAINED WITH US-BASED OR MRI-BASED(PSEUDO-) GROUND-TRUTH

Bayesian CNNs were created applying MC dropout at test-time to the CNN networks trained with the US-based ground-truths and the MRI-based pseudo-ground-truths (Sections II.D.2a and II.D.2b). The number of passes for each image at test-time was fixed to 20.

For comparison, the CNN networks trained with the US-based ground-truths and the MRI-based pseudo-ground-truths (Sections II.D.2a and II.D.2b) were also tested with no dropout at test-time, as performed in standard segmentation tasks. This type of solution will be referred to as to ''baseline'' throughout the next sections of the paper.

E. EVALUATION METRICS

The metrics reported in the following subsections were utilised to compare the baseline against the Bayesian CNN performance (Section II.D.4). Regardless of the type of training (US-based or MRI-based), the Area Under The Curve (AUC)-ROC curves (Section II.E.1) and the Dice Score Coefficient (DSC) (Section II.E.2) were obtained with respect to both the US-based ground-truths and the MRI-based pseudo-ground-truths.

For the Bayesian CNNs, the metrics below were evaluated at different uncertainty levels, similarly to Nair *et al.* [19]. The metrics were first computed considering all the predictions (as performed for the baseline), regardless of the corresponding uncertainty level (or variance). In the plots/tables in Section III, this case will be referred to as ''Variance <= 0.25'', where 0.25 was the maximum variance level in the whole dataset. Then, the metrics were re-computed retaining only the predictions with the corresponding variance (or uncertainty) below or equal to 0.15 and 0.05, respectively. Hence, in these evaluations, only a certain percentage of pixels was considered, which is reported in the plots/tables in Section III with the corresponding variance level.

1) AUC-ROC CURVES

Receiver operating characteristics (ROC) curves [27] were generated plotting the true positive rate (or sensitivity Eq. 5) against the false detection rate (or (1 - specificity) Eq. 6), computed binarising the predictions at different threshold levels between 0 and 1.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{5}$$

where TP and FN indicate the true positives and false negatives, respectively.

$$Specificity = \frac{TN}{(TN + FP)} \tag{6}$$

where TN and FP indicate the true negatives and false positives, respectively.

The AUC was also computed for each of the plotted curves.

2) DICE SCORE COEFFICIENT

The DSC [28] (Eq. 7) was computed to measure the overlap between the ground-truth M_{GT} and the prediction M_P obtained from the CNN:

$$DSC = \frac{2(M_{GT} \cdot M_P)}{|M_{GT}| + |M_P|} \tag{7}$$

where $\cdot$ represents the dot product, and $|M_{GT}|$ and $|M_P|$ are the number of positive elements in each of the masks.

Alternatively, the DSC can also be expressed in terms of true positives (TP), false positives (FP) and false negatives (FN) (Eq. 8) as:

$$DSC = \frac{2TP}{(2TP + FP + FN)} \tag{8}$$

3) DICE SCORE COEFFICIENT WITH BOUNDARY UNCERTAINTY

The Dice Score Coefficient with Boundary Uncertainty (DSC_{BU}) [8] (Eq. 9 and 12) was also computed to measure the model performance comparing the model prediction M_P with a probabilistic ground-truth $M_{GT_{BU}}$, which combined the US-based ground-truth and the MRI-based pseudo-ground-truth:

$$DSC_{BU} = \frac{2(M_{GT_{BU}} \cdot M_P)}{|M_{GT_{BU}}| + |M_P|} \tag{9}$$

To generate the probabilistic ground-truth $M_{GT_{BU}}$, one should associate to each ground-truth pixel its uncertainty level. For simplicity, in this article, we defined two uncertainty levels,

low or high uncertainty, and split the corresponding ground-truth pixels into two sets: the internal ground-truth I_{GT} and the uncertainty margin UM_{GT}. The I_{GT} included the pixels associated with low uncertainty, i.e. pixels certainly belonging to the cartilage; while the UM_{GT} included the pixels associated with high uncertainty, i.e. pixels for which we are less confident that could be part of the cartilage. The I_{GT} was defined as the union of the overlapping pixels of the US-based ground-truth ($M_{GT_{US}}$) and the MRI-based pseudo-ground-truth ($M_{PGT_{MRI}}$) (Eq. 10); the uncertainty margin UM_{GT} as the pixels either belonging to the US ground-truth or the MRI-based pseudo-ground-truth (non-overlapping) (Eq. 11).

$$I_{GT} = M_{GT_{US}} \odot M_{PGT_{MRI}} \tag{10}$$

$$UM_{GT} = \left(M_{GT_{US}} + M_{PGT_{MRI}}\right) - I_{GT} \tag{11}$$

where $\odot$ represents the Hadamard product (elementwise multiplication).

The probabilistic ground-truth was then expressed as the combination of the internal ground-truth (I_{GT}) and the overlap between the uncertainty margin (UM_{GT}) and the prediction M_P (Eq. 12).

$$M_{GT_{BU}} = I_{GT} + UM_{GT} \odot M_P \tag{12}$$

The latter term in Eq 12 indicated that since the pixels in the UM_{GT} were possibly part of the cartilage they should be considered as correct whenever included in the prediction, but also that the prediction should not be penalised for the undetected pixels of the UM_{GT}. The UM_{GT} can be considered as the region where the boundary of the structure can physically be, because it is inherently impossible to determine its precise position from the US images alone and thus it is uncertain whether the pixels in this area belong to the cartilage. Figure 4 shows an example of US image overlaid with the US-based ground-truth ($M_{GT_{US}}$) and the MRI-based pseudo-ground-truth ($M_{PGT_{MRI}}$) (Figure 4.a) and the corresponding I_{GT}, UM_{GT} (Figure 4.b).

F. REPRODUCIBILITY OF MRI-US REGISTRATION

The intra-observer consistency in performing the registration was calculated, since the MRI-based pseudo-ground-truth was created based on the MRI-US manual registration performed by the expert. The surgeon repeated the MRI-US registration (Section II.B.2.b) for 6 volume pairs in two separate sessions. The norm of the difference vector of the translations (d) and rotations (r) values selected by the surgeon for the MRI volume during each registration session [24] were calculated using the following equations (Eqs. 13-14).

$$d = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \tag{13}$$

$$r = \sqrt{\Delta \alpha^2 + \Delta \beta^2 + \Delta \gamma^2} \tag{14}$$

where: $\Delta x = x_1 - x_2$; $\Delta y = y_1 - y_2$; $\Delta z = z_1 - z_2$; $\Delta \alpha = \alpha_1 - \alpha_2$; $\Delta \beta = \beta_1 - \beta_2$; $\Delta \gamma = \gamma_1 - \gamma_2$ and x, y, z and α, β, γ are the three translation and rotation values, respectively, selected by the surgeons in either session 1 or 2 as indicated by the subscript.

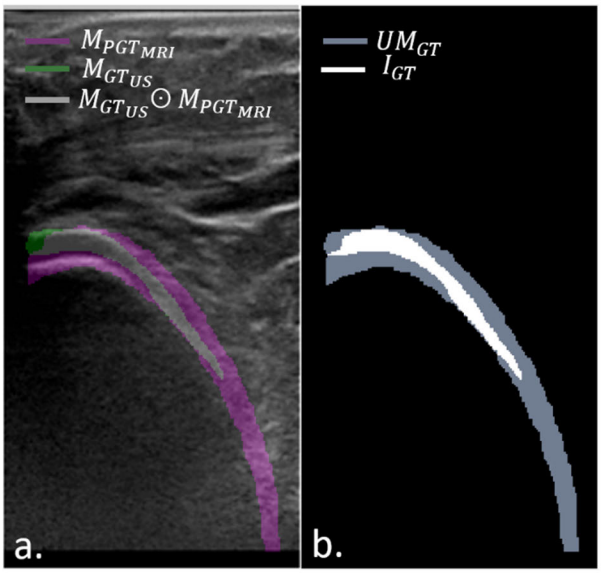

FIGURE 4. Internal ground-truth (I_{GT}) and uncertainty margin (UM_{GT}) generation from the US-based ground-truth ($M_{GT_{US}}$) and the MRI-based pseudo-ground-truth ($M_{PGT_{MRI}}$). a) US image example overlaid with the $M_{GT_{US}}$ (green) and the $M_{PGT_{MRI}}$ (purple). The overlap between the two ($M_{GT_{US}} \odot M_{PGT_{MRI}}$) is shown in white. b) The corresponding I_{GT} and UM_{GT} shown in white and grey, respectively.

Furthermore, to analyse the direct effect of the MRI registration error, for each of the 6 MRI-US pairs, the two registrations performed by the surgeon were used to generate two different MRI pseudo-ground-truths following the procedure described in Sections II.B.2b and II.B.2c. The overlap of the resulting segmentations on the US images was used to compute their discrepancy.

III. RESULTS

A. BAYESIAN CNN VS BASELINE TRAINED WITH US-BASED GROUND-TRUTH

The ROC curves indicate a performance increase when the Bayesian CNN was utilised to generate the predictions (magenta curves in Figure 5, AUC values in Table 3). This trend was confirmed both comparing the CNN prediction to the standard US-based ground-truths (Figure 5a-c) and to the MRI-based pseudo-ground-truths (Figure 5b-d). In addition to the AUC, the maximum DSC with respect to both the standard US-based ground-truths and the MRI-based pseudo-ground-truths and the maximum DSCub were also computed for Datasets 1 and 2 (Table 3). The maximum values were selected among the DSC and DSCub values calculated after binarising the predictions at thresholds levels between 0 and 1, with an incremental step of 0.001.

The Max DSC with respect to the MRI pseudo-ground-truth and the Max DSCub (highlighted in yellow in Table 3) increased by 4% to 6% and by 1% to 3%, respectively for the prediction obtained with dropout at test-time (all variance levels considered, where the maximum variance was

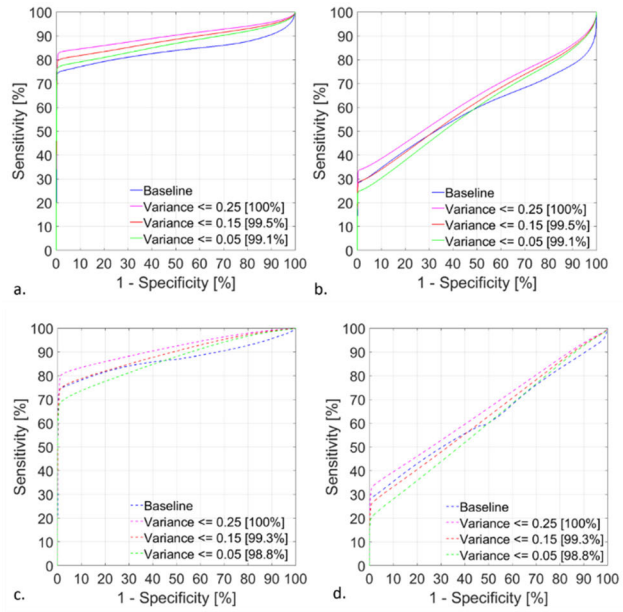

FIGURE 5. ROC curves for Dataset 1 comparing the performance of the baseline (CNN without Dropout at test-time) and the predictions generated using Dropout at test-time at different variance (uncertainty) thresholds: all uncertainty levels (0.25), 0.15 and 0.05. The retained pixels for the corresponding uncertainty levels are reported in each sub-figure legend. The ROC curves on the left side and the right side of the figure compare the different predictions to the standard US-based ground-truths and the MRI-based pseudo-ground-truths, respectively. a)-b) Correspond to a CNN training using a Dropout probability of discarding the network weights of 10%; c)-d) Correspond to a CNN training using a Dropout probability of discarding the network weights of 50%.

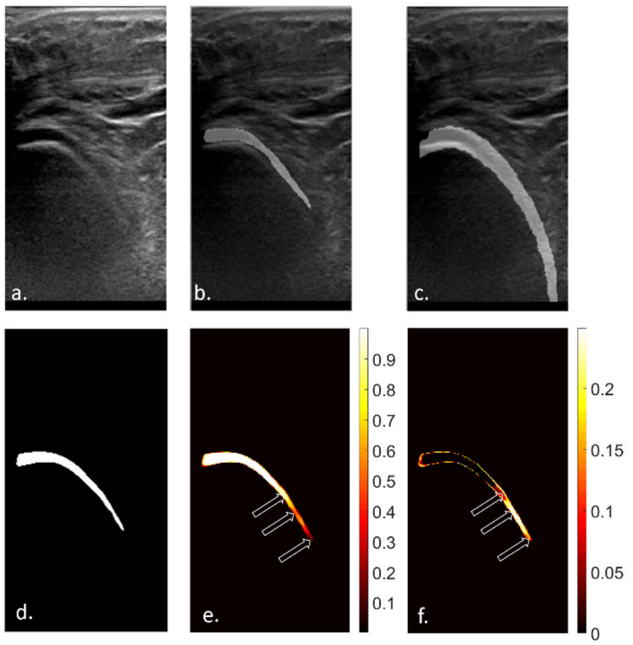

FIGURE 6. Mean prediction and variance distribution on an US image example. a) Original US image; b) Overlay of US image and corresponding standard US-based ground-truth; c) Overlay of US image and corresponding MRI-based pseudo-ground-truth; d) Baseline (UNet generated prediction); e) Mean prediction (using Dropout at test-time, p=10%); f) Variance (using Dropout at test-time, p=10%).

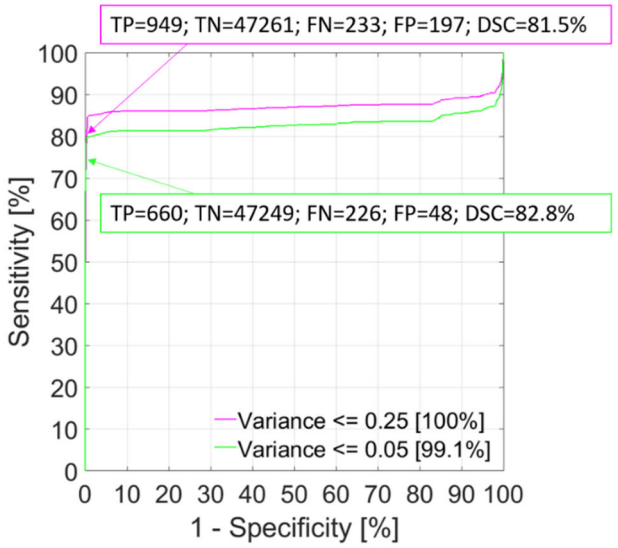

FIGURE 7. ROC curves of US image example Figure 6 comparing the predictions generated to the standard US-based ground-truths, using Dropout at test-time when all the variance levels are considered (magenta) vs pixels for variance levels equal or smaller than 0.05 were retained. The number of true positives (TP), true negatives (TN), false negatives (FN), false positives (FP) and the dice are reported for an example threshold equal to 0.1. For lower variance levels (low CNN uncertainty), the number of FN and FP (the number of wrongly classified pixels) is reduced, but also a significant number of TP is discarded, leading to lower AUC, compared to the case where all variance levels are considered.

0.25) compared to the baseline (CNN with no dropout at test-time).

When the highest levels of uncertain pixels were discarded (pixels corresponding to levels of variance higher than either 0.15 or 0.05), the AUC, the max DSC with respect to the MRI-based pseudo-ground-truth and the Max DSCub dropped compared to the case where all levels of variance were retained (red and green curves vs magenta curve in Figure 5, AUC values in Table 3). This result can be justified by the presence of the highest levels of uncertainty on relatively large areas of the US images where the cartilage region was not well defined (cartilage boundaries and inner part were rough and not homogenous). A representative example is provided in Figure 6, where the uncertainty level was higher (Figure 6f) corresponded to the area where the cartilage was not well defined (Figure 6a). Discarding the CNN prediction in these areas reduced the error rate of the prediction (the number of false positives and false negatives), but it also led to discarding a relatively large part of true positives. This behaviour is shown in Figure 7 where the ROC curves for a representative US image of the lateral-medial area of the cartilage (Figure 6) are depicted, and the corresponding number of TPs, TNs, FNs, FPs are reported for an example threshold equal to 0.1. It should also be noted that the DSC coefficient was slightly higher for the case where the variance levels higher than 0.05 were discarded, as this metric assigns a high penalty to wrongly classified pixels.

TABLE 3. AUC, maximum DSC and maximum DSCub for the baseline (CNN without Dropout at test-time) and the predictions generated using Dropout at test-time at different variance (uncertainty) thresholds: all uncertainty levels (0.25), 0.15 and 0.05. Column 1 reports the Dropout probability of discarding the network weights (10% or 50%). Column 2 reports if the prediction was assessed with the standard US-based ground-truth, with the MRI-based pseudo-ground-truth or with the probabilistic ground-truth.

			Dataset 1				Dataset 2			
Dropout p	Evaluation	Prediction type	Retained pixels [%]	AUC [%]	Max DSC [%]/ Threshold	Max DSC_{ub} [%]/ Threshold	Retained pixels [%]	AUC [%]	Max DSC [%]/ Threshold	Max DSC_{ub} [%]/ Threshold
10%	Prediction vs US GT	**Baseline***	**100**	**83.7**	**75.8[0.001]**	-	**100**	**82.7**	**74.8[0.001]**	-
		Var<=Max (0.25)	**100**	**90.0**	**76.7[0.051]**	-	**100**	**89.2**	**75.1[0.152]**	-
		Var<=0.15	99.5	88.3	76.8[0.051]	-	99.6	88.0	76.8[0.152]	-
		Var<=0.05	99.1	86.6	77.4[0.052]	-	99.2	86.7	78.5[0.103]	-
	Prediction vs MRI P-GT	**Baseline***	**100**	**57.9**	**41.5[0.001]**	-	**100**	**50.1**	**41.4[0.001]**	-
		Var<=Max(0.25)	**100**	**63.5**	**47.3[0.001]**	-	**100**	**59.1**	**45.3[0.001]**	-
		Var<=0.15	99.5	60.9	41.8[0.001]	-	99.6	56.9	41.4[0.001]	-
		Var<=0.05	99.1	58.7	37.0[0.001]	-	99.2	55.1	38.4[0.001]	-
	Prediction vs probabilistic GT	**Baseline***	**100**	-	-	**86.2[0.001]**	**100**	-	-	**84.9[0.001]**
		Var<=Max(0.25)	**100**	-	-	**89.0[0.001]**	**100**	-	-	**86.0[0.051]**
		Var<=0.15	99.5	-	-	87.6[0.001]	99.6	-	-	**86.0** [0.051]
		Var<=0.05	99.1	-	-	86.3[0.001]	99.2	-	-	86.3[0.051]
50%	Prediction vs US GT	**Baseline**	**100**	**86.9**	73.3[0.001]	-	**100**	**83.8**	76.3[0.001]	-
		Var<=Max(0.25)	**100**	**91.7**	72.4[0.101]	-	**100**	**91.2**	75.8[0.100]	-
		Var<=0.15	99.3	89.4	70.4[0.101]	-	99.4	88.9	74.9[0.100]	-
		Var<=0.05	98.8	86.7	69.8[0.065]	-	98.9	86.3	75.0[0.064]	-
	Prediction vs MRI P-GT	**Baseline**	**100**	**62.2**	**39.8[0.001]**	-	**100**	**53.3**	**41.4[0.001]**	-
		Var<=Max(0.25)	**100**	**66.4**	**44.9[0.001]**	-	**100**	**63.2**	**45.9[0.001]**	-
		Var<=0.15	99.3	63.0	37.1[0.001]	-	99.4	59.6	38.6[0.001]	-
		Var<=0.05	98.8	60.1	30.1[0.001]	-	98.9	56.6	31.8 [0.001]	-
	Prediction vs probabilistic GT	**Baseline**	**100**	-	-	**83.4[0.001]**	**100**	-	-	**86.4[0.001]**
		Var<=Max(0.25)	**100**	-	-	**84.0[0.051]**	**100**	-	-	**87.2[0.051]**
		Var<=0.15	99.3	-	-	80.4[0.051]	99.4	-	-	84.5[0.050]
		Var<=0.05	98.8	-	-	77.6[0.051]	98.9	-	-	82.2[0.050]

*Standard CNN (no Dropout at test-time)
GT = ground-truth
P-GT = pseudo-ground-truth

Analysing more in-depth Figure 6, one can intuitively understand the mean probability and the variance distributions in relation to the US image information. The cartilage area proximal to the patella was almost perpendicular to the incident US beam and thus its boundaries resulted in being well-defined. Most of the pixels in this region had a high mean probability of belonging to the cartilage (∼1) at a high confidence (variance ∼0) of the CNN (Figure 6 e-f). Moving along the cartilage curvature, the cartilage became less and less defined on the US image. This effect resulted in a mean prediction/variance distribution typically characterised by 3 regions (as indicated by the white arrows in Figure 6 e-f): the first is the area proximal to the defined region where the mean probability started to drop and the CNN started to be less confident; the second region comprises most of the area where the cartilage was not well-defined corresponding to intermediate mean probabilities levels and the highest uncertainty level of the algorithm; and a final transition region between the cartilage and the background, where the mean probability became close to 0 at high confidence (variance around 0.05-0.10).

B. BAYESIAN CNN VS BASELINE TRAINED WITH MRI-BASED PSEUDO-GROUND-TRUTH

As for the training using the standard US-based segmentation, the ROC curves comparing the CNN prediction with the two types of ground-truths (Figure 8a-d) showed a performance increase when the Bayesian CNN was used. Similarly to the previous case analysed, the Max DSC with respect to the standard US-based ground-truths and the maximum DSCub reported a performance increase with respect to the baseline of 7-8% and about 2%, respectively (highlighted in yellow in Table 4).

The ROC curves for the MRI-based pseudo-ground-truths showed similar behaviour to the ones shown in the previous section in terms of both the performance decrease when the highest variance levels were discarded and of the type of prediction/variance distribution generated on the US image

TABLE 4. AUC, maximum DSC and maximum DSCub for the baseline (CNN without Dropout at test-time) and the predictions generated using Dropout at test-time at different variance (uncertainty) thresholds: all uncertainty levels (0.25), 0.15 and 0.05. Column 1 reports the Dropout probability of discarding the network weights (10% or 50%). Column 2 reports if the prediction was assessed with the standard US-based ground-truth, with the MRI-based pseudo-ground-truth or with the probabilistic ground-truth.

Dropout p	Evaluation	Prediction type	Retained pixels [%]	AUC [%]	Max DSC [%]/ Threshold	Max DSC_{ub} [%]/ Threshold
10%	Prediction vs US GT	**Baseline***	**100**	**97.8**	**45.5[0.999]**	-
		Var <=0.25	**100**	**98.3**	**51.7[0.999]**	-
		Var <=0.15	98.5	98.4	52.3[0.999]	-
		Var <=0.05	97.3	98.5	52.8[0.999]	-
	Prediction vs MRI P-GT	**Baseline***	**100**	**91.0**	**72.4[0.001]**	-
		Var <=0.25	**100**	**92.5**	**72.5[0.211]**	-
		Var <=0.15	98.5	92.0	74.8[0.214]	-
		Var <=0.05	97.3	91.5	76.8[0.860]	-
	Prediction vs probabilistic GT	**Baseline***	**100**	-	-	**87.7[0.999]**
		Var <=0.25	**100**	-	-	**90.3[0.999]**
		Var <=0.15	98.5	-	-	90.9[0.999]
		Var <=0.05	97.3	-	-	91.5[0.999]
50%	Prediction vs US GT (p=50%)	**Baseline**	**100**	**97.6**	**45.5[0.999]**	-
		Var <=0.25	**100**	**98.4**	**53.2[0.949]**	-
		Var <=0.15	97.3	98.8	55.6[0.949]	-
		Var <=0.05	95.2	99.0	57.9[0.949]	-
	Prediction vs MRI P-GT	**Baseline**	**100**	**91.8**	**74.0[0.001]**	-
		Var <=0.25	**100**	**93.6**	**74.2[0.203]**	-
		Var <=0.15	97.3	92.4	76.2[0.203]	-
		Var <=0.05	95.2	91.0	78.3[0.114]	-
	Prediction vs probabilistic GT	**Baseline**	**100**	-	-	**88.5[0.999]**
		Var <=0.25	**100**	-	-	**90.2[0.794]**
		Var <=0.15	97.3	-	-	92.2[0.854]
		Var <=0.05	95.2	-	-	92.2[0.854]

*Standard CNN (no Dropout at test-time)
GT = ground-truth
P-GT = pseudo-ground-truth

(see Figure 9 e-f and the corresponding ROC curves in Figure 10). However, compared to the US-based ground-truth training, the highest uncertainty areas were smaller and covered more distal areas of the cartilage (with respect to the skin surface) which were not identified through the US-based ground-truths.

When the highest uncertainty levels of the CNN predictions were discarded and compared to the US-based ground-truths, the CNN performance increased (Figure 8 b-d and Table 4) indicating that the predicted areas where the algorithm had high confidence correlated with areas where the cartilage was contoured based on the US information. This result was further confirmed in Figure 11 that shows the mean predictions and variance corresponding to the different ground-truth regions and to the background. Mean predictions and variance corresponding to pixels belonging to the cartilage area considered as certain (overlap between standard US-based ground-truth and MRI-based pseudo-ground-truth) were binned at a 0.1 and 0.05 intervals, respectively, and the percentage of pixels in each bin was reported. The same procedure was repeated for mean predictions and variance corresponding to pixels belonging to the standard US-based ground-truth pixels (Figure 11 b); the MRI-based pseudo-ground-truth pixels (Figure 11 c) and background pixels (not belonging to US-based ground-truth or to the MRI-based pseudo-ground-truth) (Figure 11 d). As expected, for the area where the cartilage presence was considered certain almost all the pixels showed prediction $\sim$1 and variance $\sim$0. Since most of the US-based ground-truth pixels were contained in the MRI pseudo-ground-truth, a similar result is shown in the figure part b. When the predictions corresponding only to the MRI-based ground-truth were considered (Figure 11 c), most predictions showed high mean probability to belong to the cartilage with low variance, but about 22% of the pixels would be possibly classified as background (high confidence of low mean prediction).

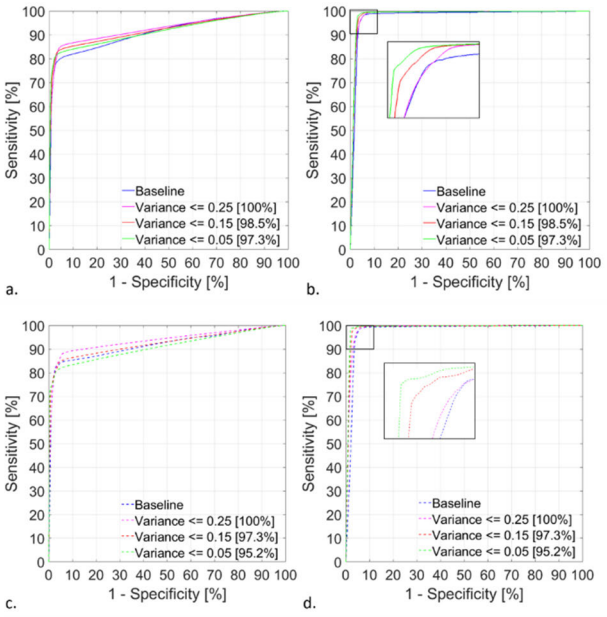

FIGURE 8. ROC curves comparing the performance of the baseline (CNN without Dropout at test-time) and of the predictions generated using Dropout at test-time at different variance (uncertainty) thresholds: all uncertainty levels (0.25), 0.15 and 0.05. The retained pixels for the corresponding uncertainty levels are reported in each sub-figure legend. The ROC curves on the left side and on the right side of the figure compare the different predictions to the MRI-based pseudo-ground-truths and to the standard US-based ground-truths, respectively. a)-b) Correspond to a CNN training using a Dropout probability of discarding the network weights of 10%; c)-d) Correspond to a CNN training using a Dropout probability of discarding the network weights of 50%.

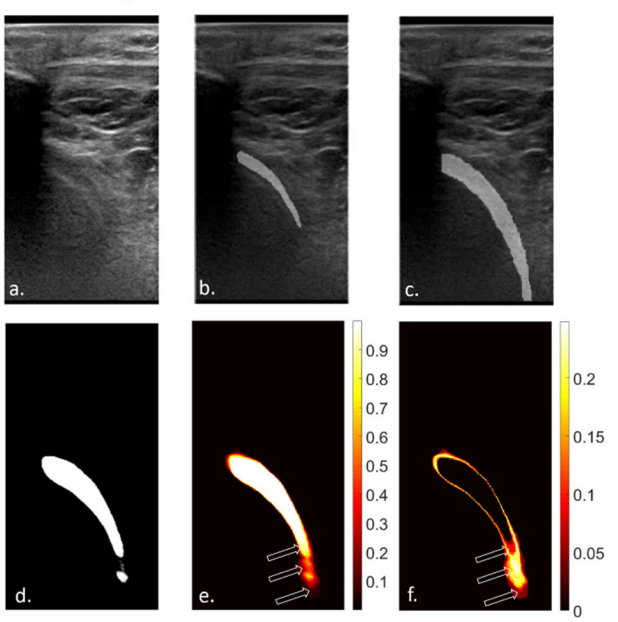

FIGURE 9. Mean prediction and variance distribution on an US image example. a) Original US image; b) Overlay of US image and corresponding standard US-based ground-truth; c) Overlay of US image and corresponding MRI-based pseudo-ground-truth; d) Baseline (UNet generated prediction); e) Mean prediction (using Dropout at test-time, p=0.1); f) Variance (using Dropout at test-time, p=0.1).

C. REPRODUCIBILITY OF MRI-US REGISTRATION

The average norm of the translation and the rotation difference vector (Section II.F) was of 1.39 mm ± 1.39 SD (range: 0 – 4.54 mm) and 3.30 degrees ± 3.23 SD (range: 0.09 – 13.53 degrees), respectively.

The discrepancy between the MRI pseudo-ground-truths generated in the different registration sessions was always located at the boundaries and varied along the cartilage curvature. It was typically lower than +/− 1 mm, for the ground-truth parts corresponding to the proximal part of the cartilage with respect to the US probe, and +/− 1- 2 mm for the more distal areas. A representative example is shown in Figure 12.

IV. DISCUSSION

The Bayesian CNN implemented in this article produced a better segmentation performance when compared against the baseline (traditional CNN), both for the training with the US-based ground-truth and for the MRI-based pseudo-ground-truth. It should be noted in particular that these results include the predictions for all the uncertainty levels provided by the algorithm. These results can be interpreted more in-depth through the DSC_{BU} and the DSC reported in this study. The former measures the overall performance since it compares the prediction with the probabilistic ground-truth, that includes the cartilage detected by both modalities.

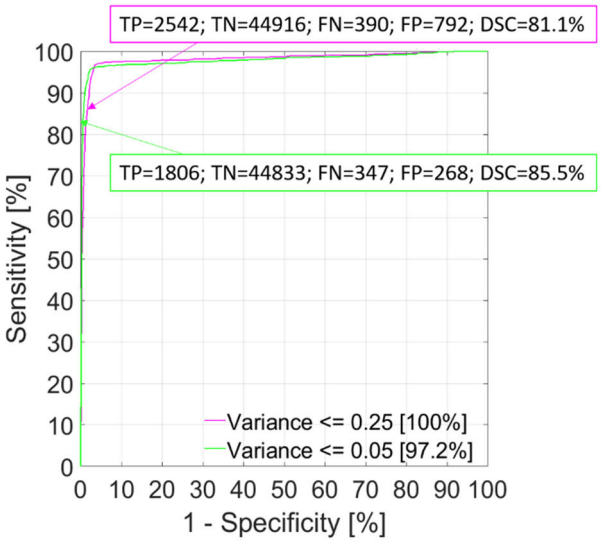

FIGURE 10. ROC curves of US image example Figure 9 comparing the predictions generated to the standard MRI-based pseudo-ground-truths, using Dropout at test-time when all the variance levels are considered (magenta) vs pixels for variance levels equal or smaller than 0.05 were retained. The number of TP, TN, FN, FP and the dice are reported for an example threshold equal to 0.1. For lower variance levels (low CNN uncertainty), the number of FN and FP (the number of wrongly classified pixels) is reduced, but also a significant number of TP is discarded, leading to lower AUC, compared to the case where all variance levels are considered.

The latter specifies the prediction performance related to each modality individually. An increase in the Max DSC_{BU} was observed for both the US and MRI-based training (by 1% to 3% and by about 2%, respectively), indicating the

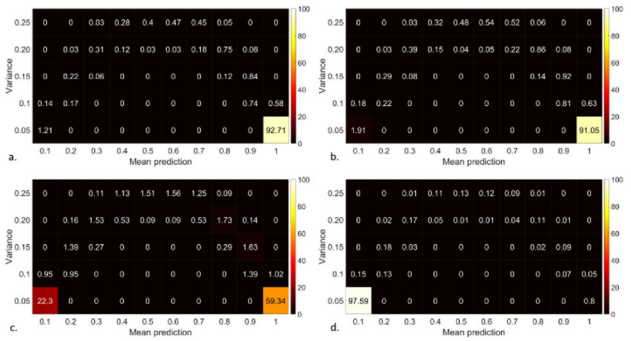

FIGURE 11. Mean prediction and variance distribution for training with Dropout probability of discarding the network weights of 10%. a) Mean predictions and variance corresponding to pixels belonging to the cartilage area considered as certain (overlap between standard US-based ground-truth and MRI-based pseudo-ground-truth) were binned at a 0.1 and a 0.05 intervals, respectively and the percentage of pixels in each bin reported. The same procedure was repeated for mean predictions and variance corresponding to pixels belonging to: b) the standard US-based ground-truth pixels; c) the MRI-based pseudo-ground-truth pixels; d) background pixels (not belonging to US-based ground-truth MRI-based pseudo-ground-truth).

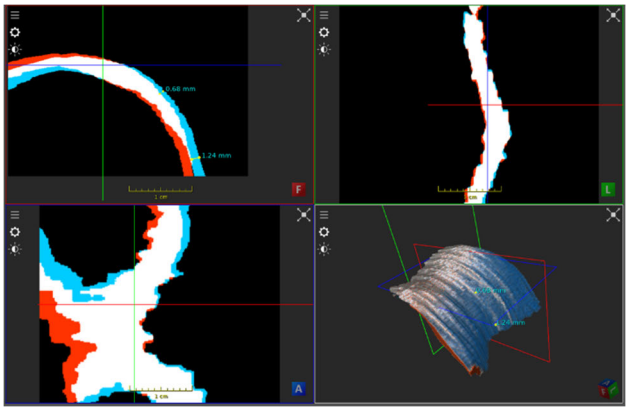

FIGURE 12. Example of overlapped MRI-based pseudo-ground-truths generated according to the corresponding MRI volume positions selected in each of the two registration sessions. The discrepancy between the two ground-truth is visualised in red and blue, while the overlapping region is shown in white.

Bayesian CNN superior ability (over the baseline) to identify the femoral cartilage correctly. More specifically, for the US-based training, this performance growth was reflected in a significant increase (by 4%-6%) of the Max DSC with respect to the MRI-based pseudo-ground-truth, while the Max DSC with respect to the US-based ground-truth did not show a significant change. This implies that the Bayesian CNN allowed identifying pixels belonging to the cartilage, as confirmed through the MRI-based pseudo-ground-truth, that were not included in the US-based ground-truth because the cartilage was not sufficiently defined to be contoured in the respective image region. For the MRI-based training, the Max DSC_{BU} performance increment was instead due to the Max DSC increase (by 7%-8%) relative to the US-based ground-truth, while the Max DSC with respect to the MRI-based pseudo-ground-truth remained almost unchanged. Thus, the Bayesian CNN was able to identify a larger part of pixels that were delineated on the US-based ground-truths but were not detected in the corresponding MRI-based pseudo-ground-truths. It should be noted that in both the US and MRI-based training, the increase in the Max DSC was higher than the DSC_{BU} as the latter does not penalise undetected pixels in the uncertainty area (i.e. pixels part of either the US-based ground-truth or the not overlapping MRI-based pseudo-ground-truth).

The MRI-based training reported in this article should be further tested on multiple volunteers. However, this approach is promising and showed several advantages over the US-based training. The segmentation performance was higher. The DSC_{BU} increased by 1%-6% and a significantly larger part of the cartilage could be detected. The AUC-ROC and the heat maps in Section III.B show the ability of the model to detect with high confidence the cartilage contoured on the US images. Furthermore, the areas where the CNN was uncertain were significantly smaller compared to the ones resulting from the US-based training. We believe the reason for this is the fact that the MRI-based labels used in training were not directly annotated on the US volumes and thus they might be more consistent in classifying the image pixels as part of the cartilage. Another advantage of this type of training is the reduced number of annotated images needed. An annotated MRI volume (about 200 MRI sagittal slices) was required to label all the US volumes of the corresponding volunteer's knee (about 120 US sagittal slices per volume), which can potentially result in thousands of images. Labels generation required additional steps compared to the US-based training, but it was still significantly more time-efficient. The main disadvantages of the MRI-based training were the need for the volunteer MRI and the possible introduction of errors when generating the pseudo-ground-truths. Label propagation from the MRI to the US volumes was based on the corresponding MRI-US manual registration. We reported a total mean error of about 1.4 mm, which translated to an error at the ground-truth boundaries of +/− 1-2 mm (Section III.C). Possible interpolation errors should also be considered since the MRI volume was resampled to the corresponding US voxel dimensions, which were an order of magnitude smaller along each direction.

Future work should aim at performing automatic feature-based MRI-US registrations to enhance the efficiency of the label propagation and possibly the registration consistency. The main challenge in the registration procedure was due to the US volume partial view of the femoral cartilage and the reduced definition of the structure along its curvature. While an accurate match between the anatomical structure in the two modalities could be found for the US regions where the cartilage was well defined, inconsistencies in the MRI-US registration may be caused by all those areas where either the cartilage was shielded by the patella or it was not possible to define its exact boundaries on the US volume.

The uncertainty approach utilised in this article could be refined by explicitly modelling the uncertainty into two separate components: aleatoric and epistemic uncertainty [29].

The aleatoric uncertainty accounts for inherent noise in the data; while the epistemic uncertainty represents the model uncertainty and thus it could be solved if enough data were to be provided to the model. The Bayesian CNN with MC dropout utilised in this article should theoretically capture the epistemic uncertainty, but it could be considered as an approximation of the two types. It has been proved that epistemic and aleatoric uncertainty are not mutually exclusive and when one of the two components is not explicitly modelled, the other one attempts to compensate for both [29]. However, while the aleatoric uncertainty is modelled during training and thus it does not interfere with the computation time at testing, the Bayesian CNN with MC sampling is time expensive. The implemented CNN can segment 125 2D US images per second (thus approximately 1 US volume per second). Utilising the Bayesian CNN, the computation time increases proportionally to the number of times each image is passed into the network for MC sampling (in this article 20 times). This is a significant limitation for an application such as surgical guidance. For the final clinical application, it would be paramount to model the aleatoric uncertainty explicitly and utilise a large training dataset to limit the epistemic uncertainty.

Previous works in this field utilising Bayesian CNN with MC dropout show a performance increase when the predictions associated with the highest levels of uncertainty were discarded and argue that uncertain predictions correlate with wrong predictions [16]–[21]. This was not necessarily the case in this study. Our results suggest that the highest levels of uncertainty corresponded to ambiguous image regions where even an expert would not be able to confidently discriminate the target from the background using the US image information only. Both qualitative and quantitative findings proved this. The highest uncertainty areas corresponded either to pixels at the boundary or to those image regions where the cartilage was present but not sufficiently defined to be contoured consistently (see Figures 6 and 9). Quantitative results show that these uncertainty areas were part of the MRI-based pseudo-ground-truth and thus part of the cartilage. A key difference with respect to other studies in the literature is the imaging modality used. Differently from CT, MRI and fundus images, where the tissue boundaries are typically well defined, US images contain many areas where the image information is not clear.

Furthermore, previous studies compared the Bayesian CNN with a deterministic ground-truth not accounting for uncertainties in the ground-truth itself. This could generate a misinterpretation of the results, where uncertain predictions may be wrongly classified as incorrect as they are compared with a clinically acceptable ground-truth that does not represent all possible clinically acceptable solutions. Segmentation tasks are in fact typically affected by relatively large intra- and inter-observer variability in the ground-truth. For this reason, we believe that as the complexity of the segmentation representation is enhanced through a probabilistic approach such as the one presented here, it is essential to consider annotation uncertainty. In this article, we showed the first attempt at assessing the Bayesian CNN with a probabilistic ground-truth utilising an additional imaging modality. An alternative solution to be explored may consist in combining labels from multiple annotators.

ACKNOWLEDGMENT

The MRI data used in this project was acquired by the Herston Imaging Research Facility, Brisbane, Australia. The authors would like to thank all the volunteers participating in this study, the students that contributed to the U.S. dataset creation and Christopher Edwards for setting the U.S. system parameters. They would like to thank Dimity Miller for sharing her knowledge about Bayesian CNNs.

REFERENCES

[1] M. Antico, F. Sasazawa, Y. Takeda, A. T. Jaiprakash, M.-L. Wille, A. K. Pandey, R. Crawford, and D. Fontanarosa, "4D ultrasound-based knee joint atlas for robotic knee arthroscopy: A feasibility study," *IEEE Access*, vol. 8, pp. 146331–146341, 2020.

[2] L. Wu, A. Jaiprakash, A. K. Pandey, D. Fontanarosa, Y. Jonmohamadi, M. Antico, M. Strydom, A. Razjigaev, F. Sasazawa, J. Roberts, and R. Crawford, "Robotic and image-guided knee arthroscopy," in *Handbook Robotic Image-Guided Surgery*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 493–514.

[3] A. Jaiprakash, W. B. O'Callaghan, S. L. Whitehouse, A. Pandey, L. Wu, J. Roberts, and R. W. Crawford, "Orthopaedic surgeon attitudes towards current limitations and the potential for robotic and technological innovation in arthroscopic surgery," *J. Orthopaedic Surgery*, vol. 25, no. 1, Jan. 2017, Art. no. 230949901668499.

[4] W. W. Curl, J. Krome, E. S. Gordon, J. Rushing, B. P. Smith, and G. G. Poehling, "Cartilage injuries: A review of 31,516 knee arthroscopies," *Arthroscopy, J. Arthroscopic Rel. Surgery*, vol. 13, no. 4, pp. 456–460, Aug. 1997.

[5] E. Smistad and L. Løvstakken, "Vessel detection in ultrasound images using deep convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10008. Cham, Switzerland: Springer, 2016, pp. 30–38.

[6] H. Ravishankar, S. M. Prabhu, V. Vaidya, and N. Singhal, "Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 779–782.

[7] A. Jaumard-Hakoun, K. Xu, P. Roussel-Ragot, G. Dreyfus, and B. Denby, "Tongue contour extraction from ultrasound images based on deep neural network," in *Proc. 18th Int. Congr. Phonetic Sci. (ICPhS)*, 2016, pp. 1–5.

[8] M. Antico, F. Sasazawa, M. Dunnhofer, S. M. Camps, A. T. Jaiprakash, A. K. Pandey, R. Crawford, G. Carneiro, and D. Fontanarosa, "Deep learning-based femoral cartilage automatic segmentation in ultrasound imaging for guidance in robotic knee arthroscopy," *Ultrasound Med. Biol.*, vol. 46, no. 2, pp. 422–435, Feb. 2020.

[9] M. Dunnhofer, M. Antico, F. Sasazawa, Y. Takeda, S. Camps, N. Martinel, C. Micheloni, G. Carneiro, and D. Fontanarosa, "Siam-U-net: Encoder-decoder Siamese network for knee cartilage tracking in ultrasound images," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101631.

[10] M. Antico, D. Vukovic, S. M. Camps, F. Sasazawa, Y. Takeda, A. T. Le, A. T. Jaiprakash, J. Roberts, R. Crawford, D. Fontanarosa, and G. Carneiro, "Deep learning for US image quality assessment based on femoral cartilage boundaries detection in autonomous knee arthroscopy," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 7, pp. 1–7, Jan. 2019.

[11] G. Kompella, M. Antico, F. Sasazawa, S. Jeevakala, K. Ram, D. Fontanarosa, A. K. Pandey, and M. Sivaprakasam, "Segmentation of femoral cartilage from knee ultrasound images using mask R-CNN," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 966–969.

[12] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," Jun. 2015, *arXiv:1506.02158*. [Online]. Available: https://arxiv.org/abs/1506.02158

[13] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2018, pp. 1–7.
[14] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3D vehicle detection," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3266–3273.
[15] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," 2015, *arXiv:1511.02680*. [Online]. Available: http://arxiv.org/abs/1511.02680
[16] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander, "Bayesian image quality transfer with CNNs: Exploring uncertainty in dMRI super-resolution," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2017.
[17] Z. Eaton-Rosen, F. Bragman, S. Bisdas, S. Ourselin, and M. J. Cardoso, "Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2018.
[18] F. J. S. Bragman, R. Tanno, Z. Eaton-Rosen, W. Li, D. J. Hawkes, S. Ourselin, D. C. Alexander, J. R. McClelland, M. J. Cardoso, "Uncertainty in multitask learning: Joint representations for probabilistic MR-only radiotherapy planning," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2018.
[19] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101557.
[20] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Sci. Rep.*, vol. 7, no. 1, pp. 1–4, Dec. 2017.
[21] O. Ozdemir, B. Woodward, and A. A. Berlin, "Propagating uncertainty in multi-stage Bayesian convolutional neural networks with application to pulmonary nodule detection," Dec. 2017, *arXiv:1712.00497*. [Online]. Available: https://arxiv.org/abs/1712.00497
[22] T. Nair, D. Precup, D. L. Arnold, and T. Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham, Switzerland: Springer, 2018.
[23] M. P. Recht, D. W. Goodwin, C. S. Winalski, and L. M. White, "MRI of articular cartilage: Revisiting current status and future directions," *Amer. J. Roentgenol.*, vol. 185, no. 4, pp. 899–914, Oct. 2005.
[24] J. Vaarkamp, "Reproducibility of interactive registration of 3D CT and MR pediatric treatment planning head images," *J. Appl. Clin. Med. Phys.*, vol. 2, no. 3, pp. 131–137, 2001.
[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2004.
[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*. [Online]. Available: https://arxiv.org/abs/1505.04597
[27] L. B. Lusted, "Signal detectability and medical decision-making," *Science*, vol. 171, no. 3977, pp. 1217–1219, Mar. 1971.
[28] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.
[29] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5574–5584.

MARIA ANTICO received the B.Eng. degree in engineering sciences from the University of Rome Tor Vergata, Italy, in 2014, and the M.Eng. degree in biomechanical engineering from the Technical University of Delft, The Netherlands, in 2016. She is currently pursuing the Ph.D. degree with the Queensland University of Technology, Australia. Her research interest includes advanced tissue recognition techniques for fully automated robotic surgery.

FUMIO SASAZAWA received the degree from the Faculty of Engineering, The University of Tokyo, Tokyo, Japan, in 1997, the degree from the School of Medicine, Shinshu University, Matsumoto, Japan, to obtain medical license, in 2004, and the Ph.D. degree in cellular and molecular biology from the Graduate School of Medicine, Hokkaido University, in 2014. He has worked as a Visiting Researcher with the Medical Robotics Team, Queensland University of Technology, Brisbane, Australia, from 2017 to 2018. He is currently an Orthopaedic Surgeon specializing in lower extremities, including hip and knee joint.

YU TAKEDA received the Ph.D. degree from the Hyogo College of Medicine, Nishinomiya, Japan, in 2018. He studied medicine at the Hyogo College of Medicine, from 2003 to 2009. During the past year, he has worked as a Researcher with the Queensland University of Technology, Australia, in the field of ultrasound-guided autonomous surgery robotic applications. He is currently working as an Orthopaedic Surgeon with the Department of Orthopedic Surgery, Hyogo College of Medicine.

ANJALI TUMKUR JAIPRAKASH received the B.S. degree in biotechnology from Bangalore University, India, in 2005, and the M.S. degree in biotechnology and business and the Ph.D. degree from the Queensland University of Technology, Australia, in 2007 and 2014, respectively. She is currently working as the Advance Queensland Research Fellow of the Queensland University of Technology. Her work merges different disciplines such as medicine, engineering and design, to develop medical devices that translate robotic vision into affordable systems that can be used to improve healthcare outcomes. She has experience in the fields of medical robotics, medical devices, and orthopaedics.

MARIE-LUISE WILLE (Member, IEEE) was born in Germany, in 1983. She received the B.Sc. and M.Sc. degrees in physics from the University of Basel, Switzerland, in 2008, and the Ph.D. degree in medical physics from the Queensland University of Technology (QUT), Brisbane, Australia, in 2015. From 2009 to 2011, she was a Research Assistant with the Shock Waves Laboratory, Fraunhofer Institute for Short Time Dynamics, Freiburg, Germany. Since 2015, she has been a Postdoctoral Research Fellow with the Institute of Health and Biomedical Innovation, QUT. Since 2019, she has been the Deputy Director of the ARC Training Centre in Multiscale 3D Imaging, Modelling, and Manufacturing, QUT. Her research interests include the interdisciplinary area of biomedical engineering and medical physics applying multiscale 3D imaging and modeling to medical and non-medical problems. She is an active member of the IEEE QLD Section Committee; and was the Chair, from 2016 to 2018, and is currently the Vice-Chair, since 2019, of the IEEE Women in Engineering Affinity Group QLD Section Committee. Since 2020, she has been a member of the R10 Professional Activities Committee.

AJAY K. PANDEY is currently a Senior Lecturer in robotics and autonomous systems with the School of Electrical Engineering and Robotics and the Domain Leader of manufacturing with advanced materials with the Institute of Future Environments. Prior to his current appointment, he held prestigious fellowships, including the QUT-Vice Chancellor's Senior Research Fellowship and the Australian Renewable Energy Agency (ARENA) Research Fellowship at the University of Queensland (UQ). His research interests include the interdisciplinary mix of photonics, chemical physics, molecular electronics, computer science, and robotics. He has published widely around issues surrounding energy, environment, and health.

ROSS CRAWFORD holds the position of the Chair in orthopaedic research and the Director of the Medical Engineering Research Facility, Queensland University of Technology, Australia. With more than 200 publications, collaborations with industry and hospitals, he is an internationally recognized Expert in orthopaedics and robotic surgery.

GUSTAVO CARNEIRO received the Ph.D. degree in computer science from the University of Toronto, in 2004. In 2005, he was a Postdoctoral Fellow with The University of British Columbia and the University of California at San Diego. From 2006 to 2008, he was the Research Scientist of Siemens Corporate Research, Princeton, USA. From 2008 to 2011, he was a Marie Curie IIF Fellow and a Visiting Assistant Professor with the Instituto Superior Tecnico, Lisbon, Portugal, within the Carnegie Mellon University-Portugal Program (CMU-Portugal). In 2011, he joined The University of Adelaide as a Senior Lecturer, where he became an Associate Professor in 2015 and a Professor in 2019. In 2014 and 2019, he joined the Technical University of Munich as a Visiting Professor and a Humboldt Fellow, respectively. He is currently a Professor with the School of Computer Science, The University of Adelaide, an ARC Future Fellow, and the Director of medical machine learning with the Australian Institute of Machine Learning. His main research interests include computer vision, medical image analysis, and machine learning.

DAVIDE FONTANAROSA is currently a Physicist with a solid background in ultrasound imaging and medical physics. He worked in one of the top institutions for radiation therapy (MAASTRO Clinic, The Netherlands) and in one of the largest industrial research laboratories in the world, Philips Research, as the Senior Scientist. Then he moved to the Queensland University of Technology, Brisbane, Australia, where he is also doing research in several fields related to ultrasound, imaging techniques, and radiation therapy. He is also an Associate Professor.