Multi-Scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution CT images

Heutink, F.; Koch, V.; Verbist, B.; Woude, W.J. van der; Mylanus, E.; Huinck, W.; ... ; Caballo, M.

Computer Methods and Programs in Biomedicine 191 (2020) 105387

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Multi-Scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution CT images

Floris Heutink[a], Valentin Koch[b], Berit Verbist[c], Willem Jan van der Woude[b], Emmanuel Mylanus[a], Wendy Huinck[a], Ioannis Sechopoulos[b,d], Marco Caballo[b,1,*]

[a] Department of Otorhinolaryngology and Donders Institute for Brain, Cognition and Behavior, Radboudumc, Geert Grooteplein Zuid 10, 6525 GA, Nijmegen, the Netherlands
[b] Department of Radiology and Nuclear Medicine, Radboudumc, Geert Grooteplein Zuid 10, 6525 GA, Nijmegen, the Netherlands
[c] Department of Radiology, Leiden University Medical Centre, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands
[d] Dutch Expert Center for Screening (LRCB), Wijchenseweg 101, 6538 SW, Nijmegen, the Netherlands

ARTICLE INFO

Article history:
Received 7 November 2019
Revised 7 February 2020
Accepted 11 February 2020

Keywords:
Cochlea
Convolutional neural networks
Deep learning
Image segmentation
Ultra-high-resolution CT

ABSTRACT

Background and objective: Performing patient-specific, pre-operative cochlea CT-based measurements could be helpful to positively affect the outcome of cochlear surgery in terms of intracochlear trauma and loss of residual hearing. Therefore, we propose a method to automatically segment and measure the human cochlea in clinical ultra-high-resolution (UHR) CT images, and investigate differences in cochlea size for personalized implant planning.

Methods: 123 temporal bone CT scans were acquired with two UHR-CT scanners, and used to develop and validate a deep learning-based system for automated cochlea segmentation and measurement. The segmentation algorithm is composed of two major steps (detection and pixel-wise classification) in cascade, and aims at combining the results of a multi-scale computer-aided detection scheme with a U-Net-like architecture for pixelwise classification. The segmentation results were used as an input to the measurement algorithm, which provides automatic cochlear measurements (volume, basal diameter, and cochlear duct length (CDL)) through the combined use of convolutional neural networks and thinning algorithms. Automatic segmentation was validated against manual annotation, by the means of Dice similarity, Boundary-F1 (BF) score, and maximum and average Hausdorff distances, while measurement errors were calculated between the automatic results and the corresponding manually obtained ground truth on a per-patient basis. Finally, the developed system was used to investigate the differences in cochlea size within our patient cohort, to relate the measurement errors to the actual variation in cochlear size across different patients.

Results: Automatic segmentation resulted in a Dice of 0.90 ± 0.03, BF score of 0.95 ± 0.03, and maximum and average Hausdorff distance of 3.05 ± 0.39 and 0.32 ± 0.07 against manual annotation. Automatic cochlear measurements resulted in errors of 8.4% (volume), 5.5% (CDL), 7.8% (basal diameter). The cochlea size varied broadly, ranging between 0.10 and 0.28 ml (volume), 1.3 and 2.5 mm (basal diameter), and 27.7 and 40.1 mm (CDL).

Conclusions: The proposed algorithm could successfully segment and analyze the cochlea on UHR-CT images, resulting in accurate measurements of cochlear anatomy. Given the wide variation in cochlear size found in our patient cohort, it may find application as a pre-operative tool in cochlear implant surgery, potentially helping elaborate personalized treatment strategies based on patient-specific, image-based anatomical measurements.

1. Introduction

A cochlear implant (CI) is a surgically implanted electronic device that provides a sense of sound to a patient with severe to profound hearing loss. To date, large variability exists in preserva-

* Corresponding author.
E-mail address: marco.caballo@radboudumc.nl (M. Caballo).
[1] Institutional Address: Department of Radiology and Nuclear Medicine (Route 767), Radboudumc, Geert Grooteplein Zuid 10, 6525 GA, Nijmegen, the Netherlands.

https://doi.org/10.1016/j.cmpb.2020.105387
0169-2607/

tion of residual hearing and speech understanding abilities after cochlear implantation [1–2]. Among other factors, a major cause of residual hearing loss (RHL) is traumatic electrode insertion [3]. The potential occurrence of intracochlear trauma during electrode insertion may be related to the fact that most current electrodes are chosen and inserted independently from each specific patient's inner ear anatomy [4]. Since, currently, electrodes have a fixed size and a standard insertion length, patients with smaller cochlea may be at higher risk of trauma and, potentially, of a larger RHL.

If accurate image-based segmentation and measurements of the cochlea could be performed pre-operatively, this could potentially allow for the adaptation of size, shape and depth of insertion of the CI electrode to each single patient, possibly improving the surgical outcome by reducing risk of intracochlear trauma and RHL.

In medical images, simple measurements of regular anatomical parts are usually performed manually. However, for highly complex and irregular structures (such as the human cochlea), dedicated computerized methods are needed to first segment the structure of interest, and then provide automatic measurements which would otherwise be challenging (if not impossible) to perform by human readers.

To address the goal of segmenting and measuring the human cochlea, previous studies proposed semiautomatic segmentation methods [5] that require a high degree of human interaction to separate the cochlea from the connected internal auditory canal and vestibular structures. Other studies aimed at using segmentation frameworks based on anatomical information of the inner ear, obtained a priori using mathematical modeling or some high-resolution, high-dose cadaver scans [6–11].

These previously developed methods reported high segmentation performance, thanks to their considerable computational and mathematical complexity. However, most methods were developed based on micro-CT scans of a few cadaveric human cochleae, which were used as constraints and as *a priori* information to guide the segmentation model. The extrapolation of information from small datasets could potentially limit the application of such methods in the clinical realm, and potentially account for limited inter-patient variability and validation.

With the advancements in medical imaging technology and analysis algorithms, new solutions can be investigated, thanks both to the improved spatial resolution of the most recently developed CT scanners, and by replacing traditional model-based segmentation with new deep learning approaches.

From the imaging side, advances in computed tomography technology have been proposed over the past few years, with wider detectors being introduced [12], novel electronics with lower noise being designed [13] and, more recently, smaller detector elements being developed [14]. In this respect, ultra-high-resolution (UHR) CT (Aquilion Proteus and Precision, Canon Medical Systems Corporation) was brought to market, with a detector element size of 0.25 mm at isocenter, and with an MTF twice as high as that of current-generation multi-detector CT systems [15].

From the image analysis perspective, deep learning has become one of the major methodologies used for analyzing medical images, including image segmentation. When a sufficiently large training set is available, deep learning demonstrated high performance in the segmentation of structures with large inter-patient anatomical variability [16–17], with limited programming effort, given the ability of learning the segmentation task directly and automatically from the images, i.e. without user-selectable parameters to be tuned in a testing phase.

Among deep learning algorithms, convolutional neural networks (CNNs) have repeatedly demonstrated their high performance in many computer vision tasks [18–23], often outperforming traditional methods based on deterministic or handcrafted approaches [24]. However, they carry the drawback of the dataset size, which has to be large enough for the network to learn sufficient patterns in the input images to correctly replicate them in an independent testing phase [25]. This can be a critical issue in tomographic, high-resolution cochlea imaging, where the dataset sizes are usually limited and, therefore, can potentially limit the application of state-of-the-art Artificial Intelligence techniques.

To address this issue, in this study we developed and validated a deep learning system for cochlea segmentation and measurements that takes advantage of both extensive data augmentation, and a modular structure that localizes the segmentation task in small image regions around the cochlea, allowing to achieve good performance using a small training set. After validation, the developed system was used on clinical ultra-high-resolution (UHR) CT patient images for cochlear measurement extraction, to investigate the differences in cochlea size within a large patient cohort and relate the measurement errors to the actual variation in cochlear size across different patients.

2. Materials and methods

The proposed approach is composed of two main blocks: one for cochlea segmentation, followed by one for cochlea measurements. The cochlea segmentation block combines a computer-aided detection system based on multiscale residual CNNs with an encoder-decoder network for pixel-wise classification. The former aims at localizing the cochlea on the input temporal bone CT scans, to reduce the search space of the subsequent pixel-wise classification model, while the latter is aimed at providing automatic cochlea segmentation. The models were trained on 2D image patches extracted with a sliding-window-based approach from the cochlea scans of the training set (as explained in Sections 2.C and 2.D), and then applied in a region-based fashion on the full test set scans (as described in Section 2.E). After segmentation, the cochlea measurement block provides automatic measurements from the segmented cochlea by using a combination of CNNs and morphological thinning algorithms. All main steps of the proposed method are reported in Fig. 1.

All steps of the pipeline, along with data collection and preparation, are described in the following sections. Finally, patient-based cochlear size measurements are performed using the proposed approach on all scans of the dataset, and are analyzed to investigate the differences in cochlea size within our patient cohort.

2.A. Image acquisition

123 UHR-CT temporal bone scans were acquired and used to develop our algorithm. Images were acquired with one of two UHR-CT scanners (Aquilion Proteus and Precision, Canon Medical Systems, Otawara, Japan), both composed of a 160 multi-row detector, with an effective detector element size of 0.25 mm × 0.25 mm at the isocenter, 1792 detector channels, and a nominal focal spot of 0.4 mm × 0.5 mm (Precision) and 0.6 mm × 0.6 mm (Proteus) [26]. For this study, helical acquisitions were acquired using tube voltages of 140 kVp (Precision) and 135 kVp (Proteus), exposure time of 1.5 s, and tube currents of 100 mA (Precision) and 80 mA (Proteus), with a gantry rotation time of up to 0.35 s, and a pitch factor 0.569. The $CTDI_{vol}$ was approximately 31 mGy, measured with a 16 cm phantom (140 kVp, 150 mAs).

CT scans were performed by a trained radiographer over a cross-section of the patient head of approximately 4 cm (including the whole inner ear anatomy). That is, only a 4 cm-thick cross-section of the patient head was imaged (along the craniocaudal direction), so as to reduce the exposure by avoiding to deliver radiation dose in other regions of the patient head.

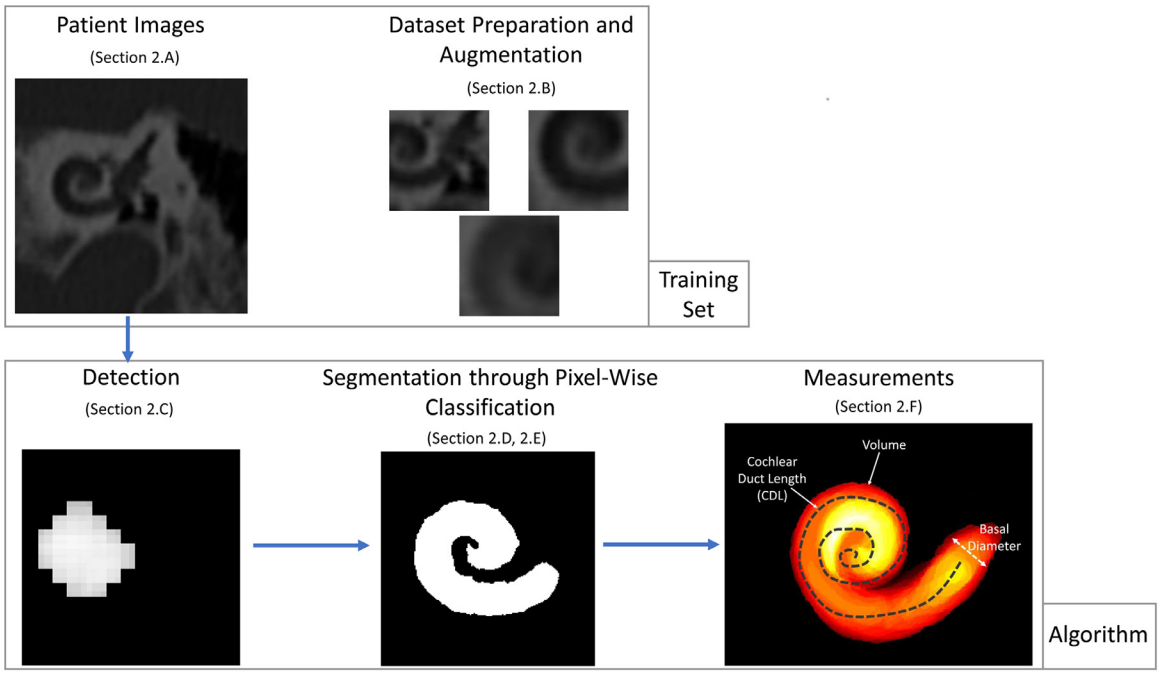

Fig. 1. Main steps of the proposed cochlea segmentation and analysis approach. The input scans are first processed by a detection module to localize the cochlea, and by a pixel-wise classification module for segmentation. The detection module aims at reducing the search space of the pixel-wise classification module, serving as pre-processing to speed up the algorithm and for false positive reduction. Both modules were trained on an image patch-basis, to increase the training set size by obtaining multiple examples from each scan. The segmented cochlear structure then undergoes a final module to extract patient-based anatomical measurements through the combination of deep learning and thinning algorithms. Deep learning was adopted in each step for its ability of learning directly from the input data, and provide automatic results without user-selectable parameters to be tuned in a testing phase.

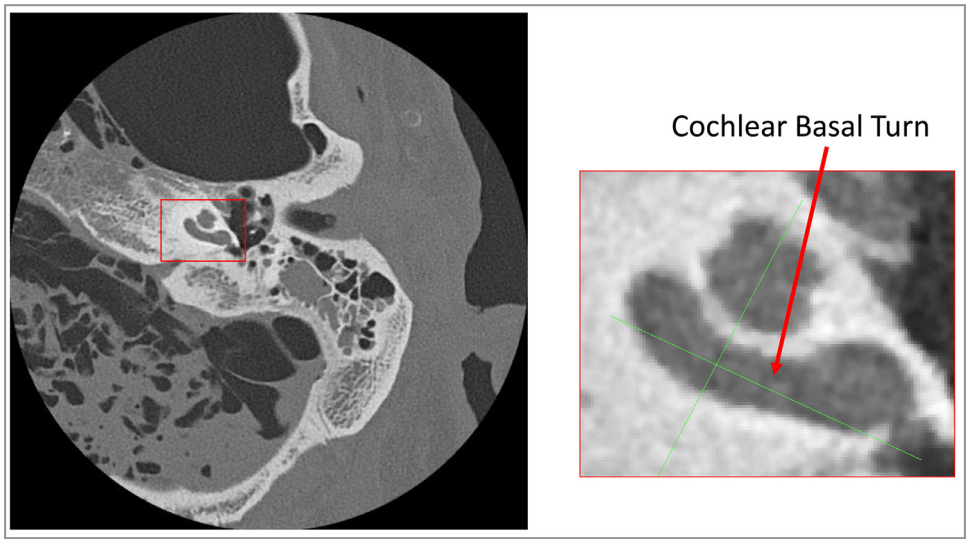

Fig. 2. Example of a temporal bone scan (axial view) showing the cochlear basal turn. Crosshairs are aligned parallel and perpendicular to the long axis of the cochlear basal turn.

The scans were then reconstructed using filtered back projection with the reconstruction kernel FC81 (a high-resolution bone kernel) along image planes parallel to the cochlear basal turn (oblique multi-plane reconstruction), with a matrix size of 1024×1024 and a slice thickness of 0.25 mm. An example of the cochlear basal turn in the axial view is shown in Fig. 2. The in-plane reconstructed voxel size was 0.045 mm for the Proteus scanner, and 0.05 mm for the Precision scanner. The reconstructed, in-plane voxel size was set automatically by the system, based on the reconstruction mode.

The dataset was collected within a prospective, cross-sectional study conducted between December 2016 and January 2018 at our institution, and approved by the local and regional medical ethics committee Arnhem-Nijmegen (METC; NL510071.091.14).

All participants of the study (average age: 64 ± 12 years for males ($n = 59$), and 61 ± 14 years for females ($n = 64$)) agreed to participate and signed informed consent. Adult patients that had undergone CI surgery between January 2010 and July 2016, after being diagnosed with post-lingual hearing loss onset (defined as an onset of severe-to-profound deafness after the age of 5 years),

were eligible for this study. The inclusion criteria were patients who could provide written informed consent, and with at least one year of experience with CI after surgery. Exclusion criteria were (i) cognitive dysfunction, and (ii) congenital or acquired anomalies of vestibulo-cochlear system.

For all patients, the cochlear structure on the image was manually annotated, and used to develop and validate our algorithms. Manual annotation was performed slice-by-slice in the reconstructed images using the ImageJ (LOCI, University of Wisconsin, NIH) polyline toolbox by a medical image analysis scientist with 3 years of experience in analysis and segmentation of CT images, under the supervision of a cochlear implant surgeon and a board-certified head-and-neck radiologist.

2.B. Data preparation and augmentation

Of the acquired scans, 40 were used to train our models, 8 for validation, while the remaining 75 were kept into an independent test set. Extensive data augmentation was performed on the training scans, in order to maximize the performance by reducing risk of overfitting while keeping most scans for final testing.

Both developed models (detection and pixel-wise classification) were trained on a patch basis. Patches were collected through a sliding window approach from each scan (and respective manual annotation) within a volume of 512 × 512 × 50 voxels (approximately corresponding to 2.5 × 2.5 × 1.25 cm) including the whole cochlear anatomy. For each cochlea scan, patches were collected in two dimensions on a slice-basis. The allowed overlay of contiguous patches was kept high (stride 10 voxels) to increase the dataset size, and the process was repeated for three different squared window sizes: 150, 100, and 70 voxel side. This multi-scale patch extraction was performed to capture the image information at different dimensions, approximately spanning from the full length of the cochlea, to the size of smaller details such as the different cochlea turns and the cochlear apex. Additional data augmentation was then performed on all extracted patches through four rotations (−20°, −10°, 10°, 20°) and vertical mirroring. These augmentation methods (and their respective parameter values) were chosen to simulate potential realistic variations in image acquisition, while avoiding generating training examples that are too different from real cases. In fact, with the imaging protocol adopted, the cochlea is always imaged at approximately the same in-plane angular orientation for all patients, justifying the use of a limited angular range (between −20° and 20°) to generate new, realistic cases.

As a result, the number of collected patches was 326,940 (150 × 150 window), 556,600 (100 × 100 window), and 904,120 (70 × 70 window). Some examples are shown in Fig. 3.

2.C. Cochlea detection model

Before segmentation, a detection model was implemented to localize the cochlear structure within the CT scan. This model outputs a probability map of the same size as the input image, with values close to 1 in those image locations where the cochlea is more likely to be present. As a detection task, it was trained with pairs of examples composed by the input image patch, and a discrete label indicating whether that patch contained cochlea voxels (1) or not (0). A training patch was assigned a label of 1 if at least a given percentage of the respective manually annotated region was composed of cochlea voxels. Given that, for a detection task, bigger field of views are more sensitive, but less specific, we set these percentages to 10%, 20%, and 25% for the 150, 100, and 70-voxel patches, respectively.

The model is composed of three residual CNNs [27–28], trained separately for the three image patch sizes, with blocks of 3 × 3 convolutional kernels plus batch normalization. The number of filters increases with the network depth (as shown in Fig. 4), and dropout regularization [29] (probability 0.5) was used before the fully connected layer for regularization. All the weights of the network were normally initialized [30], and biases set to zero. Training was performed on mini-batches [31] of 64 elements using gradient descent with a momentum of 0.9 and a weight decay of 10^{-4}. The starting learning rate was set to 0.01, and decayed exponentially every 10 epochs (over a maximum of 60 epochs). During training, accuracy was calculated on the validation set to prevent overfitting, and the loss function used was binary cross-entropy:

$$Loss = -[y \log(p) + (1-y)\log(1-p)] \tag{1}$$

where y is the ground truth label, and p the predicted detection probability.

In a testing phase, all image patches are collected from the CT scan using a sliding-window approach (10 voxels stride); the model performs the detection task for the three image patches separately, and then provides a final probability map by averaging the three outputs (Fig. 5). This results in a multi-scale probability map that evaluates the image information in a high-to-low level fashion. High-level information is restricted by the two CNNs with smaller patches, allowing to keep the sensitivity high (large patch) while devoting the specificity to the CNNs with smaller receptive fields.

2.D. Cochlea pixel-wise classification model

A second model was developed for cochlea segmentation through pixel-wise classification. For this, a U-net-like architecture [32–34] composed of an encoder-decoder structure was implemented. This model learns the segmentation task in a supervised manner, by performing a pixel-wise mapping between the original image and the manually annotated mask.

It is composed of an encoder-decoder structure as shown in Fig. 6; the encoder part vectorizes the input bidimensional feature space via 3 × 3 convolutions and max pooling [35] operations (kernel size 2 × 2, stride equal to 2 voxels), while the decoder part recovers the information via 2 × 2 nearest-neighbour up-sampling followed by two 3 × 3 convolutional kernels. The outputs of the convolutional blocks from the encoding architecture are concatenated with each corresponding decoding step, leading to a high detail preservation of the original input image. In the last layer, a 1 × 1 convolution followed by a sigmoid activation function outputs the segmentation result in the form of a pixel-wise probability.

The network was trained on the largest image patches (150×150 voxels) using mini-batches of 4 examples and the Adam (adaptive moment estimation) optimization method [36], an algorithm that adapts the learning rate for each network weight by using first and second moments of the gradient. The initial learning rate was set to 10^{-3}, with an exponential decay every 10 epochs (over a maximum of 50 epochs). The energy function was computed by a pixel-wise softmax (Eq. (2)) over the final feature map combined with the cross-entropy loss function (Eq. (3)):

$$p_i(\boldsymbol{x}) = \frac{e^{x_i}}{\sum_{j=1}^{2} e^{x_j}} \tag{2}$$

$$Loss = -\sum_{i=1}^{2} t_i \log(p_i(\boldsymbol{x})) \tag{3}$$

In Eq. (2), the non-normalized output of the network is mapped to a probability distribution over the predicted output class, where the network output is encoded by the activation values x_i of each pixel i, resulting in the pixelwise network prediction $p_i(\boldsymbol{x})$.

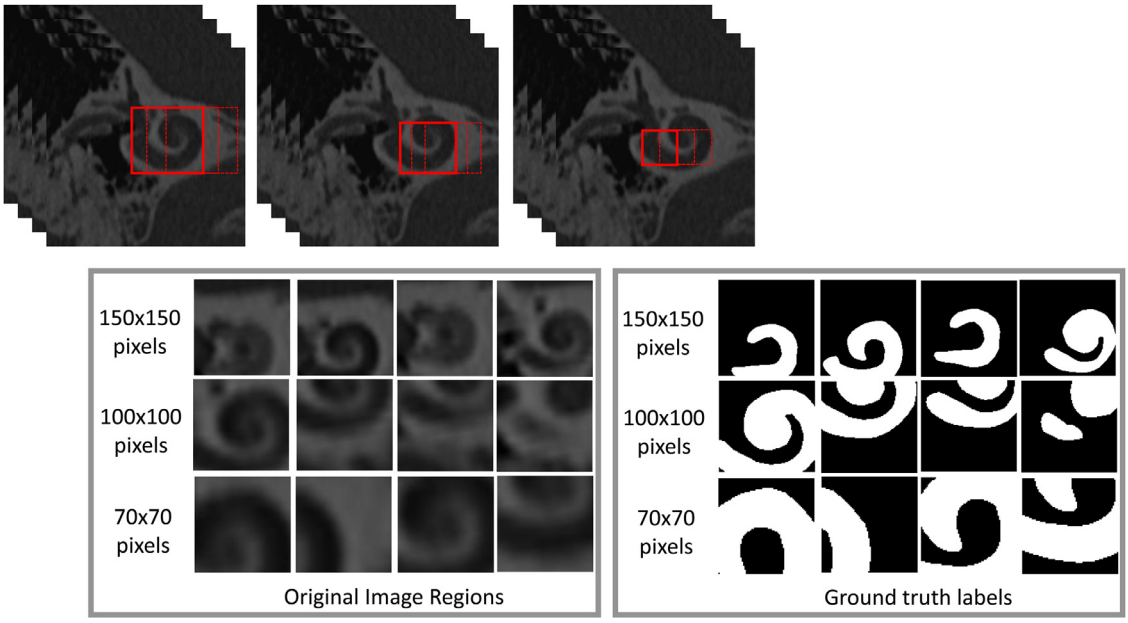

Fig. 3. Examples of training image patches for the proposed deep learning system. The patches were extracted through a sliding window approach, with the window size varying for three different sizes (150, 100, and 70 voxel side). The patches were extracted on a 2D-basis from each slice of the reconstructed cochlea scan, resulting in approximately the same number of patches extracted from each scan.

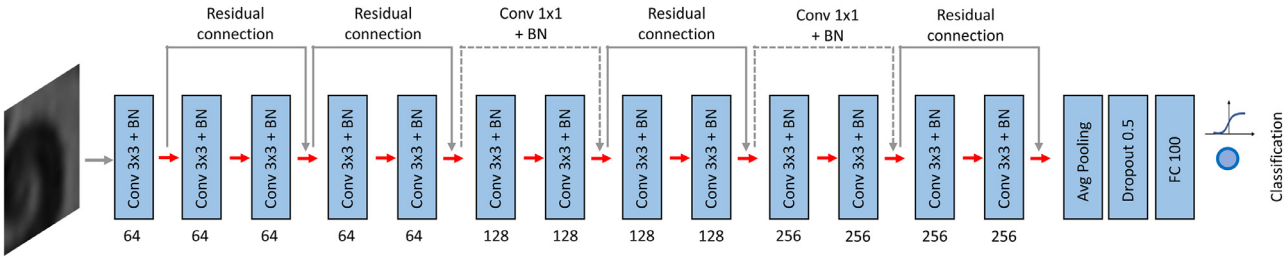

Fig. 4. Residual network architecture used for the cochlea detection model.

In Eq. (3), the learning of the network is performed by penalizing (i.e. increasing) the loss in case of wrong predictions (compared to the ground truth labels t_i).

As for the detection model, accuracy during training was calculated on the validation set to prevent overfitting.

2.E. Main algorithm for cochlea segmentation

The whole algorithm for cochlea detection and segmentation (Fig. 7.a) combines the two previously described models (detection and pixel-wise classification).

The algorithm requires a single starting seed point to be defined at any location within the part of the image occupied by the cochlear volume. Given that the cochlea (or part of it) is always located approximately in the central area of each scan, this point was selected as the central pixel of each image. After this initialization, a square window (150 × 150 voxels) is generated around the seed, and processed by the detection model. A probability score is assigned to the window, with a probability higher than 0.5 associated with a positively predicted outcome. Then, an 8-connected macro-region is grown starting from the seed point, with the macro-region containing 8 squared regions obtained by radially translating the first window along 8 different directions (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°), with a stride of 10 voxels. Translations of 45°, 135°, 225°, 315° were obtained by moving the region of interest by ±10 voxels in each direction in the (x,y) plane. These regions, translated diagonally, in addition to those translated by 0°, 90°, 180°, and 270°, were included to increase the algorithm sensitivity in this first step.

Each region is processed by the detection model, and the output probability of overlaying region parts is averaged. The process is iterated from each new positively predicted region, resulting in the macro-region to grow and cover the whole cochlear structure, and stops when no further areas fulfilling the detection criterion (probability ≥0.5) are found. The process is repeated for contiguous slices (always starting from the same seed point location) until positive regions are found, resulting in a 3D local probability map which displays the probability of voxels being cochlea.

The same process is then applied for the other two window sizes (100 and 70 voxels side), and a final map is generated by averaging the three outputs.

After this step, the pixel-wise classification model is applied in the same manner as for the detection model (for partially overlaying regions, logical *or* operation is applied), and the result is then multiplied on a voxel-by-voxel level with the final probability map derived from the detection model. Voxels are then rounded to obtain the final segmentation.

2.F. Algorithm for automatic cochlea measurements

After segmentation, the extracted cochlear structure undergoes an automatic analysis (Fig. 7.b) to compute three different measurements: volume, cochlear duct length (CDL), and basal diameter of the cochlear basal turn. Cochlear volume is automatically mea-

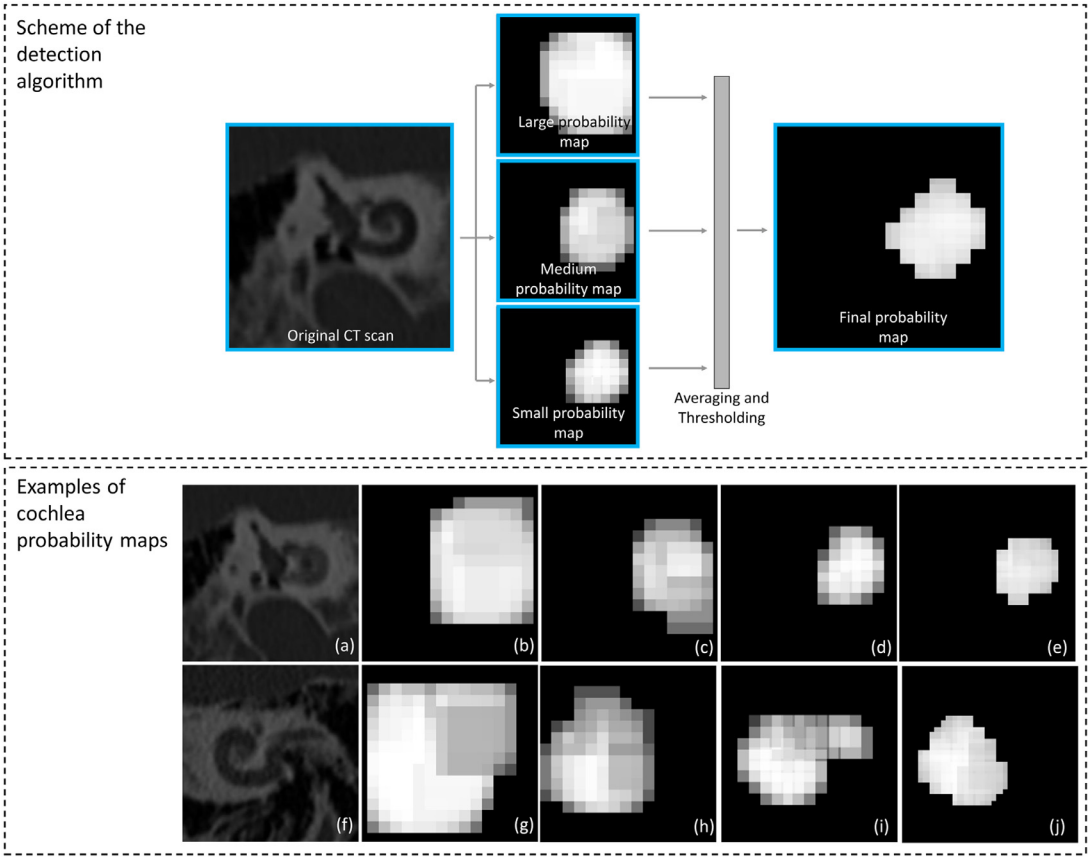

Fig. 5. (top) Schematic representation of the detection algorithm, which outputs a final probability map by averaging the results of the three CNNs (window size 150, 100, and 70 voxel side, stride equal to 10 voxels); (bottom) two examples of automatic cochlea detection showing all generated probability maps (b, g: 150 × 150 window; c, h: 100 × 100 window; d, i: 70 × 70 window; e, j: final map). The two examples show how the cochlea detection task can benefit from the proposed multi-scale approach. Especially, the second example shows how false positives (i.e. the connected auditory canal incorrectly detected by the 70 voxel-side CNN, panel (i)) are reduced and corrected in the final probability mask (panel (j)).

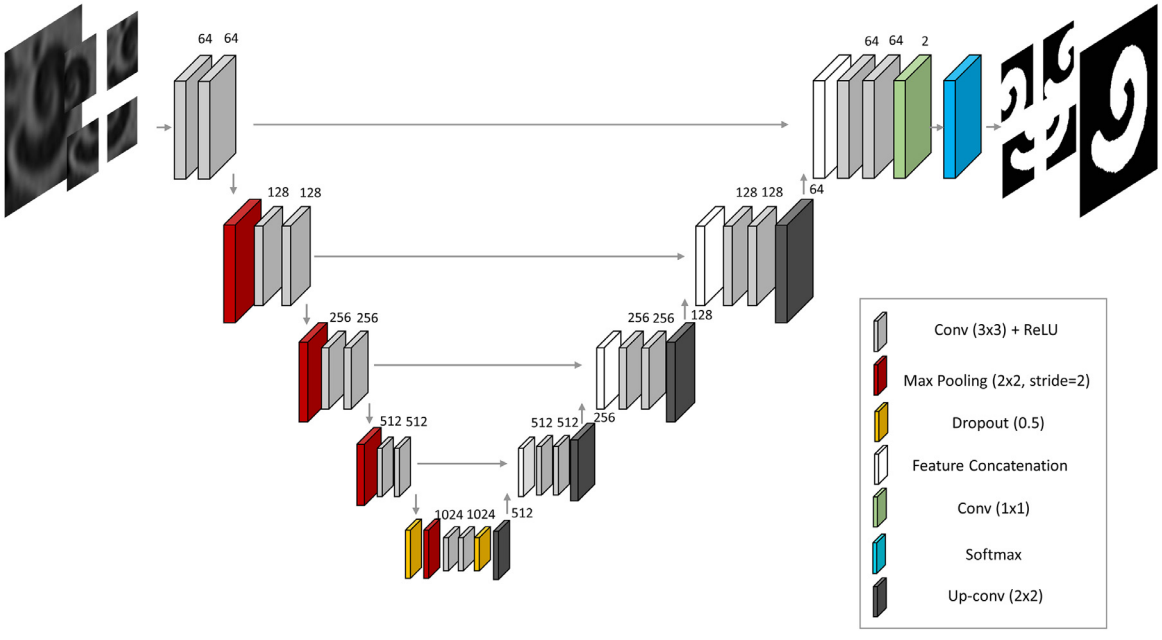

Fig. 6. Encoder-decoder network used in the pixel-wise classification model.

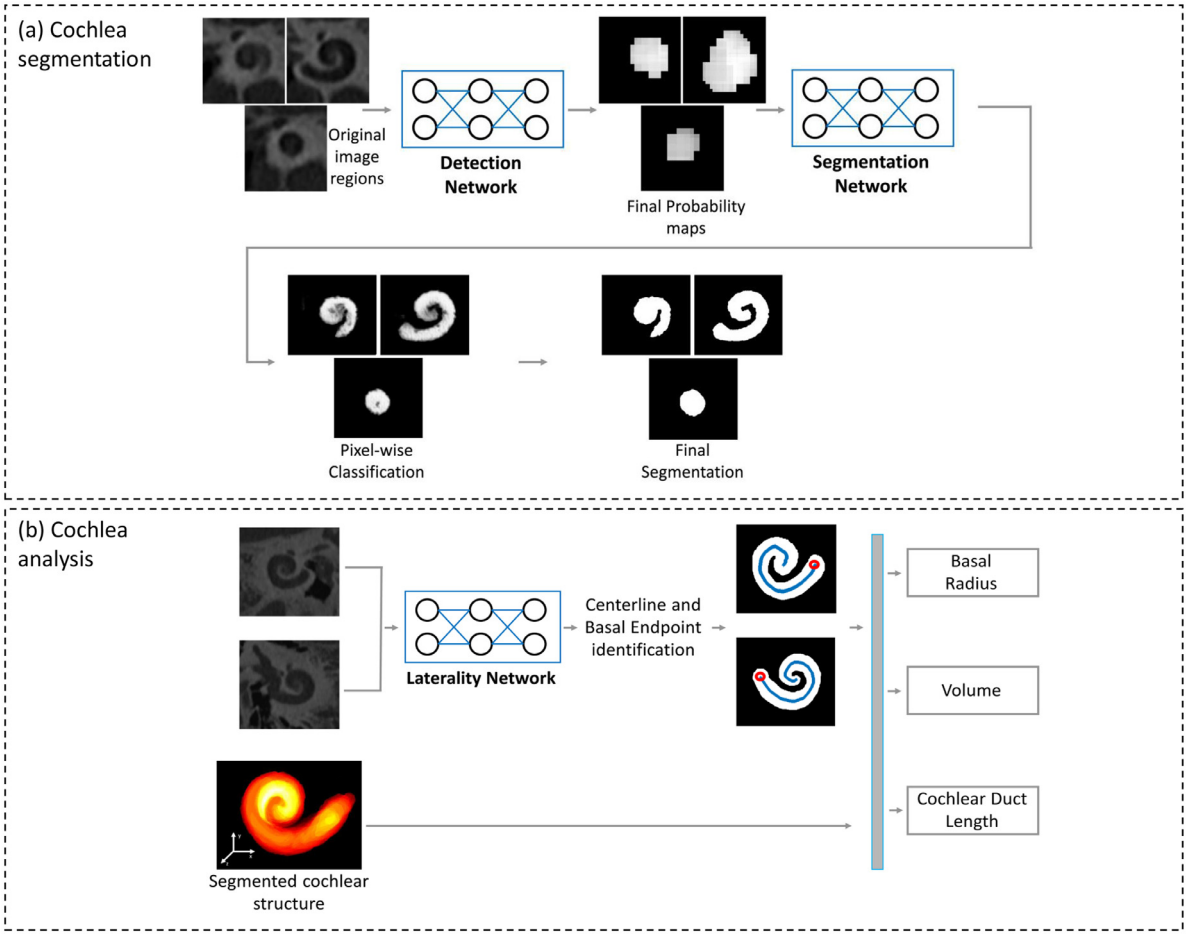

Fig. 7. Schemes for the (a) segmentation and (b) analysis algorithm.

sured by counting the number of cochlea voxels identified by the segmentation. The CDL was calculated from a previously proposed equation, which reliably estimates the distance from the middle of the round window to the helicotrema starting from the cochlear length [37], obtained as the longest dimension of the 3D bounding box enclosing the segmented cochlea. To automatically measure the basal diameter, the centerline was extracted from the 2D image slice containing the largest amount of cochlea voxels through a well-established iterative thinning algorithm [38], and the diameter was calculated as twice the distance between the basal endpoint of the centerline and the outer cochlear wall. The basal endpoint was automatically identified and distinguished from the apex endpoint through an additional CNN, which was simply trained to recognize the inner ear laterality given an input CT image slice. The CNN has the same architecture and hyperparameters as the ones used for the detection model, and was trained on 1163 examples (427 left, 536 right cochleae), where each example was a single CT slice displaying the full inner ear anatomy. The network was then tested on an additional 200 slices (100 left, 100 right) and achieved 100% accuracy.

2.G. Algorithm performance evaluation

The proposed pipeline was tested on 75 cochlea scans. The results of automatic cochlea segmentation were compared with the manually annotated scans using the four following metrics [39–41].

- Dice similarity, which measures the intersection between the two samples A and B over their union, ranging between 0 (no overlap) and 1 (perfect overlap)

$$Dice = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{4}$$

- Boundary F1 (BF) score: defined as the harmonic mean of the precision (P) and sensitivity (S), it measures how close the boundary of the segmented object matches the ground truth contour

$$BF\ score = 2 \cdot \frac{P \cdot S}{P + S} \tag{5}$$

- Hausdorff distance, which calculates the highest distance (d) between the contours of the two compared samples (the lower is the value, the better is the segmentation result)

$$HD = max\left\{ \begin{matrix} max_{a \in A} min_{b \in B}[d(a,b)], \\ max_{b \in B} min_{a \in A}[d(a,b)] \end{matrix} \right\} \tag{6}$$

- Averaged Hausdorff distance, which replaces the maximum operator in previous equation by averaging all distances, resulting in a more robust metric less sensitive to outliers.

For the automatic cochlea measurements, basal diameter, volume, and CDL were validated against the same measurements performed on the manually annotated ground truth scans. All ground truth measurements were compared with the automatic measurements on a per patient basis.

After validation, automatic measurements were extracted from all patient scans, to investigate the differences in cochlea size in the full cohort.

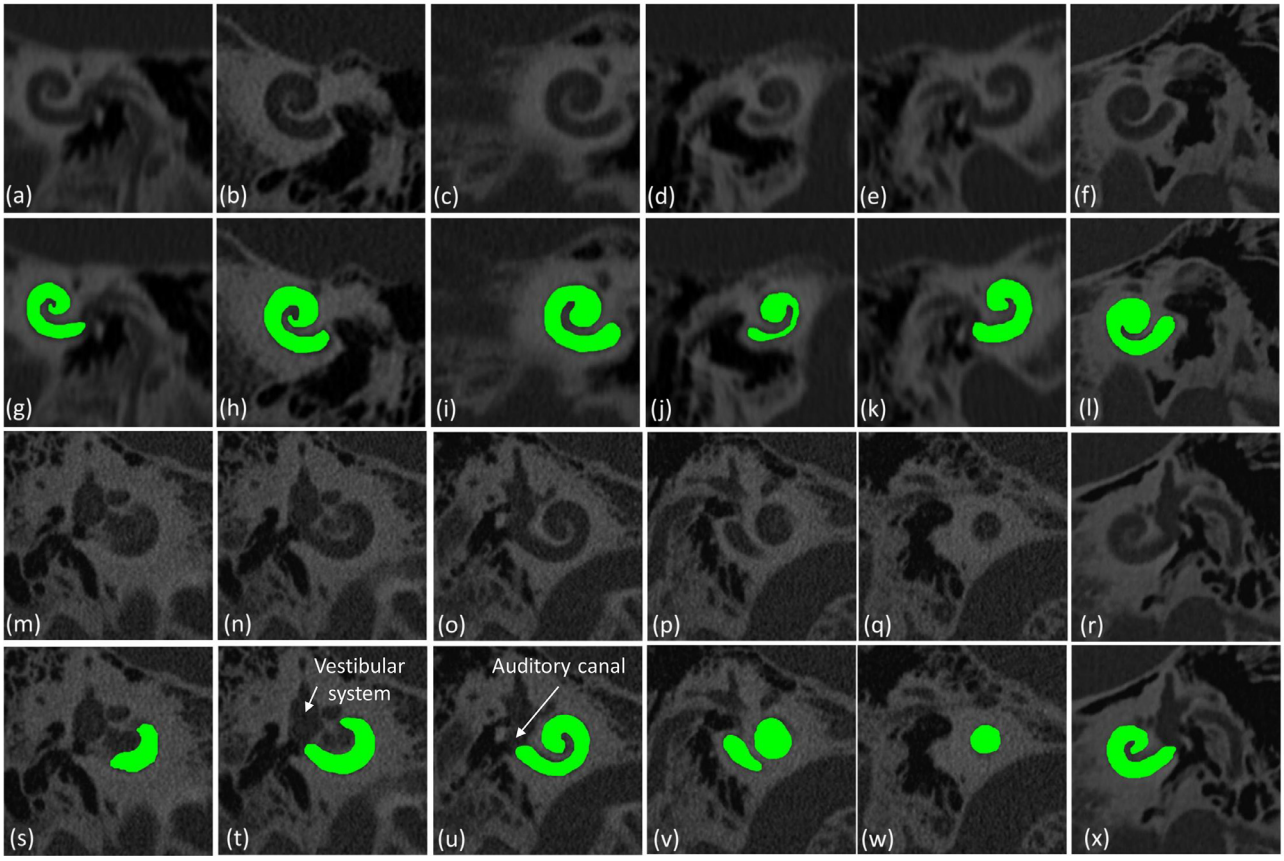

Fig. 8. Examples of original cochlea image slices (a-f, m-r), and respective segmentation results (g-l, s-x). The algorithm could correctly avoid the other structures connected to the cochlea, especially the vestibular system (t) and the external auditory canal (s, t, u). Images in panels (a), (c)–(f), (r) were acquired with the Precision UHR-CT scanner, while images in panels (b), (m)–(q) were acquired by the Proteus UHR-CT scanner.

Table 1
Results of automatic cochlea segmentation, compared to ground truth manual annotation.

Dice		BF-Score	Max Hausdorff Distance (voxel)	Average Hausdorff Distance (voxel)
Mean	0.90	0.95	3.05	0.32
Maximum	0.99	0.99	4.03	0.52
Minimum	0.87	0.92	1.62	0.11
Median	0.89	0.94	3.10	0.33
Standard deviation	0.03	0.02	0.39	0.07

Table 2
Errors between automatic and manual cochlear measurements.

	Volume Error [ml]	Basal Diameter Error [mm]	CDL Error [mm]
Mean	0.013	0.134	1.693
Maximum	0.036	0.429	4.182
Minimum	0.0004	0	0.007
Median	0.011	0.115	1.657
Standard deviation	0.008	0.100	1.130

3. Results

Some examples of automatic segmentation are reported in Fig. 8.

The testing of the segmentation algorithm (Table 1) against the manually annotated ground truth resulted in a Dice of 0.90 ± 0.03, BF score of 0.95 ± 0.03, Hausdorff distance of 3.05 ± 0.39 voxels, and averaged Hausdorff distance of 0.32 ± 0.07 voxels.

Automatic measurements (Table 2) resulted in absolute errors of 0.01 ml ± 0.008 ml (8.4%, volume), 1.69 mm ± 1.1 mm (5.5%, CDL), and 0.13 mm ± 0.10 mm (7.8%, basal diameter).

The size of the cochlea varied broadly among the patients in our dataset, ranging between 0.10 and 0.28 ml (volume), 1.3 and 2.5 mm (basal diameter), and 27.7 and 40.1 mm (CDL) (Fig. 9 and Table 3).

4. Discussion

In this study, a deep learning-based system capable of segmenting and measuring the human cochlea on UHR-CT images was developed and validated, and used to investigate the differences in cochlea size in a large patient cohort.

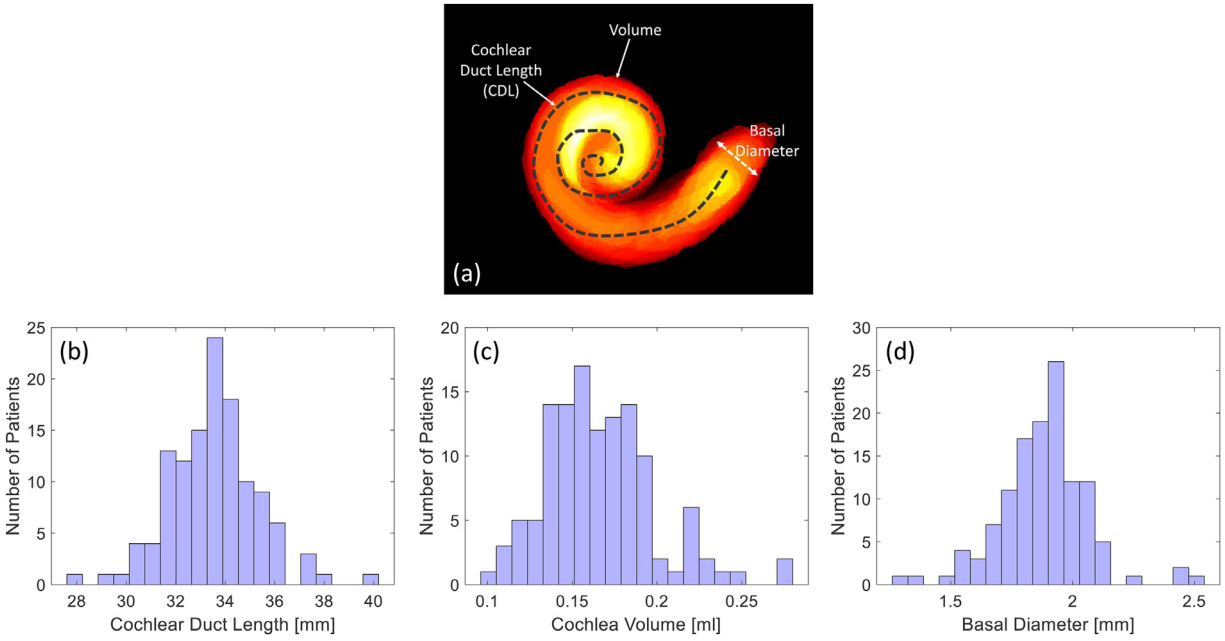

Fig. 9. (a) Scheme of the extracted cochlear size measurements; (b)–(d) Histograms of the cochlear measurements for the evaluated patient cohort.

Table 3
Mean, maximum, minimum, median and standard deviation values for the cochlea measurements, for all patients in our cohort.

Volume [ml]		Basal Diameter [mm]	Cochlear Duct Length [mm]
Mean	0.165	1.879	33.519
Maximum	0.280	2.531	40.127
Minimum	0.100	1.299	27.727
Median	0.161	1.877	33.620
Standard deviation	0.031	0.184	1.805
Mean			
Male ($n = 59$)	0.171 ± 0.031	1.912 ± 0.184	33.815 ± 1.884
Female ($n = 64$)	0.161 ± 0.030	1.849 ± 0.181	33.246 ± 1.697

The proposed system resulted in accurate cochlea segmentation and measurements, thanks to an extensive training data augmentation and the combination of different deep learning techniques.

The region-based, modular approach presented in this work achieved high performance by overcoming the issue of a limited dataset size (due to UHR-CT being made available in clinics only very recently). In fact, as opposed to full image-based approaches (that usually require much larger training sets), working on a multi-scale patch basis allowed to obtain a massive dataset from few scans, helping avoid the common issues related to deep learning and dataset size (such as overfitting and class imbalance) and leave the majority of the images for independent testing.

Incorporating a detection model before the encoder-decoder network allowed to reduce the search space of the segmentation network, decreasing the risk of false positives (examples shown in Fig. 5) and the computational time (about 10 min for segmenting and analyzing a full cochlea on a 2.7 GHz CPU, 8GB RAM workstation). Furthermore, the multi-scale approach for cochlea detection helped keep the specificity high, by localizing the segmentation only around cochlea regions, and therefore by avoiding anatomical parts which may be fully connected to the cochlea and that present the same intensity, such as the internal auditory canal, blood vessels or the vestibular system (Fig. 8), without losing in sensitivity thanks to the averaging process of different window sizes. Finally, restricting the segmentation only around regions showing the object of interest helped reduce the feature space processed by the subsequently applied U-Net, making the training process of the pixel-wise classification model easier and less prone to overfitting as opposed to working on a full image-basis.

The segmentation algorithm resulted in an average error of 10% (Dice), and the automatic measurements resulted in the highest error for the volume measurement (8%). Although these errors could limit the accuracy of our methods when measuring the human cochlea, their impact is implicitly lowered by the large variability in cochlea size across different patients highlighted by our results. This strengthens the reliability of the proposed approach in detecting the differences in cochlea size among different patients, holding the potential of being incorporated, in future and after additional extensive validation, into the cochlear imaging pipeline as a decision-making tool for cochlear implant surgery.

The main limitation of the proposed methods is the difficulty to objectively validate the measurements extracted from the segmented cochlea. For all comparisons, we considered manual annotation and measurements as the ground truth, due to the sparsity (or unavailability) of other validation methods. A more accurate approach could be performed using measurements with high-resolution, high-dose micro-CT scans acquired from cadavers, and comparing these results with the ones extracted from the same cochlea acquired in a clinical setting. However, a limited number of cadaver images is available, which would limit the validation process to too few cases. Furthermore, it would still be a completely image-based validation process, therefore potentially biased by specific image characteristics. This could be solved if the measurements were physically performed on cochlea samples, but this

approach carries the additional limitations of a very low number of available specimens, along with the difficulty to accurately drill the surrounding temporal bone for sample preparation.

Prior to moving to personalized cochlear implant modeling, several technical issues need to be overcome. Specifically, before implementing the proposed measurement approach on a clinical routine-basis, prospective clinical trials need to be performed to investigate whether the size of the cochlea correlates with the surgical and clinical outcome of cochlear implantation, currently performed with fixed-size electrodes. Especially, correlation between cochlear size and post-operative loss of residual hearing should be investigated, to test the hypothesis that smaller cochleae could be at higher risk of traumatic electrode insertion which, in turn, would lead to a higher loss residual hearing. Furthermore, since UHR-CT is not yet commonly used in clinical practice, the possibility of obtaining similar cochlear measurement performance with conventional CT should be investigated. While previous studies showed encouraging results in cochlear segmentation obtained from conventional CT [6–8], much larger datasets are needed for testing. Since such datasets are currently unavailable, we can only hypothesize that UHR-CT helps reduce measurement errors compared to conventional CT (given the small size of the human cochlea), and consequently potentially leads to an improved surgical outcome in personalized cochlear implant surgery. Therefore, while UHR-CT and Artificial Intelligence seem to have promising applications for personalized surgical planning, future studies are needed to confirm their effective performance, quantify their effect on the surgical outcome, and evaluate the potential advantages over normal resolution CT.

In addition to this, future work includes the collection of additional patient scans to further assess the appropriateness of our methods, and potentially the development of other computational strategies to further improve the segmentation performance (for example, 3D-based methods that take advantage of weakly-supervised learning to address the issue of annotating a large dataset). Finally, the extracted cochlea measurements will be related to the loss of residual hearing after cochlear implant surgery, to investigate the effect of cochlear size on speech recognition abilities after cochlear implantation.

5. Conclusions

The developed computerized system was successfully applied to extract automatic and accurate cochlear measurements based on UHR-CT images, thanks to the combination of multiple deep learning approaches and extensive data augmentation. The system highlighted a large variability in cochlea size in a large patient cohort, suggesting that the proposed approach could therefore potentially be useful as a pre-operative tool for future personalized cochlear implant surgery.

Declaration of Competing Interest

The authors have no relevant conflicts of interest to disclose.

Acknowledgments

Partial funding for this research has been provided by Canon Medical Systems Corporation. The study data and results were generated and controlled at all times by the research personnel at Radboudumc, with no influence from Canon.

References

[1] N.R. Peterson, D.B. Pisoni, R.T Miyamoto, Cochlear implants and spoken language processing abilities: review and assessment of the literature, Restor. Neurol. Neurosci. 28 (2) (2010) 237–250.
[2] B.R. Whiting, K.T. Bae, M.W Skinner, Cochlear implants: three-dimensional localization by means of coregistration of CT and conventional radiographs, RadiologyRadiology 221 (2001) 543–549.
[3] C.W. Turner, B.J. Gantz, C. Vidal, A. Behrens, B.A Henry, Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing, J. Acoust. Soc. Am. 115 (April (4)) (2004) 1729–1735.
[4] M.F. Dorman, P.C. Loizou, D Rainey, Simulating the effect of cochlear-implant electrode insertion depth on speech understanding, J. Acoust. Soc. Am. 102 (November (5 Pt 1)) (1997) 2993–2996.
[5] D. Xianfen, C. Siping, L. Changhong, W Yuanmei, 3D semi-automatic segmentation of the cochlea and inner ear, Conf. Proc. IEEE Eng. Med. Biol. Soc. 6 (2005) 6285–6288.
[6] J.H. Noble, B.R. Rutherford, R.F. Labadie, O. Majdani, B.M Dawant, Modeling and segmentation of intra-cochlear anatomy in conventional CT, Conf. Proc. SPIE Med. Im. 7623 (2010) 762302.
[7] J.H. Noble, R.F. Labadie, O. Majdani, B.M Dawant, Automatic segmentation of intracochlear anatomy in conventional CT, IEEE Trans. Biomed. Eng. 58 (September (9)) (2011) 2625–2632.
[8] F.A. Reda, B.M. Dawant, T.R. McRackan, R.F. Labadie, J.H Noble, Automatic segmentation of intra-cochlear anatomy in post-implantation CT, Conf. Proc. SPIE Med. Im. 8671 (2013) 867101.
[9] F.A. Reda, T.R. McRackan, R.F. Labadie, B.M. Dawant, J.H Noble, Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients, Med. Image Anal. 18 (April (3)) (2014) 605–615.
[10] E.R. Pujadas, H.M. Kjer, S. Vera, M. Ceresa, M.A.G Ballester, Cochlea segmentation using iterated random walks with shape prior, Conf. Proc. SPIE Med. Im. 9784 (2016) 97842U.
[11] B.M. Verbist, L. Ferrarini, J.J. Briaire, A. Zarowski, F. Admiraal-Behloul, H. Olofsen, J.H. Reiber, J.H Frijns, Anatomic considerations of cochlear morphology and its implications for insertion trauma in cochlear implant surgery, Otol. Neurotol. 30 (4) (2009) 471–477.
[12] F.J. Rybicki, et al., Initial evaluation of coronary images from 320-detector row computed tomography, Int. J. Cardiovasc. Imaging 24 (2008) 535–546.
[13] X. Duan, et al., Electronic noise in CT detectors: impact on image noise and artifacts, AJR 201 (2013) 626–632.
[14] A. Hata, et al., Effect of matrix size on the image quality of ultra-high-resolution ct of the lung: comparison of 512 × 201; 512, 1024 × 1024, and 2048 × 2048, Acad. Radiol. 25 (7) (2018) 869–876.
[15] L. Oostveen, et al., Physical Evaluation of an Ultra-High-Resolution CT Scanner, European Radiology, In Press, 2020, doi:10.1007/s00330-019-06635-5.
[16] M.U. Dalmış, et al., Using deep learning to segment breast and fibroglandular tissue in MRI volumes, Med. Phys. 44 (2) (2017) 533–546.
[17] W. Sun, X. Huang, T.-.L.B. Tseng, W Qian, Automatic lung nodule graph cuts segmentation with deep learning false positive reduction, SPIE Proc. Med. Im. 10134 (2017), doi:10.1117/12.2251302.
[18] Russakovsky O., et al. ImageNet large scale visual recognition challenge. arXiv:1409.0575, 2015.
[19] K. Ganesan, et al., Computer-Aided breast cancer detection using mammograms: a review, IEEE Rev. Biomed. Eng. 6 (2013) 77–97.
[20] L. Oakden-Rayner, et al., Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework, Sci. Rep. 7 (1) (2017) 1648 10.
[21] G. Litjens, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[22] M.A. Al-antari, et al., A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification, Int. J. Med. Inform. 117 (2018) 44–54.
[23] J.Z. Cheng, et al., Computer-Aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans, Sci. Rep. 15 (6) (2016) 24454.
[24] D. Truhn, et al., Radiomic versus convolutional neural networks analysis for classification of contrast-enhancing lesions at multiparametric breast MRI, RadiologyRadiology 290 (2) (2019) 290–297.
[25] T. Kooi, et al., Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.
[26] R. Tanaka, K. Yoshioka, H. Takagi, J.D. Schuijf, K Arakita, Novel developments in non-invasive imaging of peripheral arterial disease with CT: experience with state-of-the-art, ultra-high-resolution CT and subtraction imaging, Clin. Radiol. 74 (January (1)) (2019) 51–58.
[27] K. He, X. Zhang, S. Ren, J Sun, Deep residual learning for image recognition, IEEE Conf on Comp Vis and Patt Recogn, 2016 arXiv:1512.03385..
[28] J. Peng, S. Kang, Z. Ning, et al., Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging, Eur. Radiol. (2019), doi:10.1007/s00330-019-06318-1.
[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[30] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, PMLR 9 (2010) 249–256.
[31] M. Li, T. Zhang, Y. Chen, A.J Smola, Efficient mini-batch training for stochastic optimization, KDD (2014) 661–670.
[32] O. Ronneberger, P. Fischer, T Brox, U-Net: convolutional networks for biomedical image segmentation, MICCAI (2015) arXiv:1505.04597..
[33] A.D. Weston, P. Korfiatis, T.L. Kline, K.A. Philbrick, P. Kostandy, T. Sakinis, M. Sugimoto, N. Takahashi, B.J Erickson, Automated abdominal segmentation

of CT scans for body composition analysis using deep learning, Radiology 290 (March (3)) (2019) 669–679.
[34] 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, RadiologyRadiology 288 (July (1)) (2018) 177–185.
[35] J. Nagi, F. Ducatelle, G.A Di Caro, Max-pooling convolutional neural networks for vision-based hand gesture recognition, Conf. Proc. IEEE ICSIPA (2011), doi:10.1109/ICSIPA.2011.6144164.
[36] D.P. Kingma, Ba J. Adam, A method for stochastic optimization, Conf Proc Int Conf for Learning Representations, 2015 arXiv:1412.6980..
[37] G. Alexiades, A. Dhanasingh, C Jolly, Method to estimate the complete and two-turn cochlear duct length, Otol. Neurotol. 36 (June (5)) (2015) 904–907.
[38] L. Lam, S.W. Lee, C.Y Suen, Thinning methodologies-a comprehensive survey, IEEE Trans. Pattern Anal. Mach. Intell. 14 (9) (1992) 869–885.
[39] A.A. Taha, A. Hanbury, An efficient algorithm for calculating the exact Hausdorff distance, IEEE Trans. Pattern Anal. Mach. Intell. 37 (11) (2015 Nov) 2153–2163.
[40] G. Csurka, D. Larlus, F Perronnin, What is a good evaluation measure for semantic segmentation? BMVC (2013), doi:10.5244/C.27.32.
[41] M.P. Dubuisson, A.K. Jain, A modified Hausdorff distance for object matching, in: Proceedings of 12th International Conference on Pattern Recognition (ICPR '94), 1994, pp. 566–568, doi:10.1109/ICPR.1994.576361.