Western Graduate&PostdoctoralStudies

Western University

Scholarship@Western

Electronic Thesis and Dissertation Repository

4-18-2022 9:30 AM

Automated Segmentation of the Inner Ear and Round Window in Computed Tomography scans using Convolutional Neural Networks

Kyle A. Rioux, *The University of Western Ontario*

Supervisor: Ladak, Hanif M, *The University of Western Ontario*
Joint Supervisor: Agrawal, Sumit K, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Engineering Science degree in Electrical and Computer Engineering

© Kyle A. Rioux 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Biomedical Commons

Recommended Citation

Rioux, Kyle A., "Automated Segmentation of the Inner Ear and Round Window in Computed Tomography scans using Convolutional Neural Networks" (2022). *Electronic Thesis and Dissertation Repository*. 8461.
https://ir.lib.uwo.ca/etd/8461

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Computed tomography (CT) scans are acquired prior to cochlear implant (CI) surgery. Three-dimensional segmentations of the inner ear (IE) and round window (RW) based on clinical CTs can improve the CI procedure. Software pipelines are presented here which employ convolutional neural networks to automatically segment the IE and RW. The first pipeline produces high resolution segmentations of the IE and RW in tightly cropped CTs. Mean IE Dice score and RW centroid error were 0.88, 0.57mm and 0.93, 0.18mm in implanted and non-implanted samples, respectively. The second pipeline automatically segments the IE in large field of view CTs of any rotational orientation. Mean Dice scores of 0.83 and 0.89 were achieved in implanted and non-implanted samples, respectively. This is the first known study to present automated segmentations of the IE and RW in CTs with CIs. The pipelines provide quick and accurate segmentations.

Keywords: Automatic Image Segmentation, Computer Vision, Machine Learning, Convolutional Neural Network, Cochlear Implant, Inner Ear, Round Window

Summary for Lay Audience

Cochlear implants are a revolutionary invention which provides hearing to deaf people who have the most common form of hearing loss. Cochlear implants are surgically inserted into the inner ear, a tiny bony labyrinth within the temporal bone of the skull. The round window membrane is a natural opening from the middle ear to the inner ear and is the most frequently used entry point when surgically inserting a cochlear implant into the inner ear. Computed tomography scans are 3D medical images which are acquired prior to the cochlear implant procedure to give surgeons an understanding of the patient's anatomy. Structures within a computed tomography scan, such as the inner ear, can be labeled by hand or by using an algorithm; these labels represent 3D models of the structures and are referred to as segmentations. Segmentations of a patient’s inner ear and round window are useful for the following applications: surgeons can build an enhanced understanding of a patient's anatomy, surgical rehearsal and training platforms could integrate specific patient cases, drilling locations and electrode insertion angles can be determined for robot-assisted cochlear implant surgery, and the creation of customized cochlear implant frequency maps could be automated. Manual labeling of medical imaging data is not feasible for integration to clinical practice as it requires expert knowledge and is extremely time consuming. This work presents an automated approach which quickly and accurately segments the inner ear and round window based on any computed tomography scan with minimal user input. The automated segmentation approach utilizes a supervised machine learning algorithm (convolutional neural networks), and post-processing scripts which finalize the output segmentation. The automated approach was validated by use of clinical and cadaveric scans. Segmentations produced by the network were compared to manual expert segmentations as well as segmentations of high-resolution imaging acquisitions. This is the first known study to present an automated approach which segments the inner ear and round window in scans with cochlear implants. The approach presented in this study utilized a wide variety of imaging acquisitions, resulting in a robust model which has promising clinical implementations.

Co-Authorship Statement

This master's thesis is an integration of a journal article, and work to be submitted to a conference. The first article, Chapter 3, is currently submitted and under review in *Scientific Reports*. The second article, Chapter 4, will be submitted to a conference.

Chapter 3: Rioux KA, Helpard LW, Ladak HM, Agrawal SK. A deep learning solution for automated segmentation of the inner ear and round window on electrode-implanted and non-implanted computed tomography scans. Submitted to *Scientific Reports*.

Chapter 4: Rioux KA, Helpard LW, Ladak HM, Agrawal SK. A deep learning solution for automated segmentation of the inner ear on electrode-implanted and non-implanted clinical computed tomography with a large field of view and varying orientation. To be submitted to a conference.

The initial motivation for these studies was presented by S.K. Agrawal and H.M. Ladak. My contributions to the studies include writing, data preparation, data analysis, and software development. L.W. Helpard contributed with data preparation and consultation.

Acknowledgements

I would like to thank my supervisors Hanif Ladak and Sumit Agrawal for being tremendous role models throughout my time as a summer research student and a graduate student with the Auditory Biophysics Lab. Dr. Ladak's attention to detail and work ethic is absolutely unparalleled; it was immediately clear that the bar was set high, which explains the success and productivity of the lab. Dr. Ladak's effectiveness as a professor was evident by his popularity amongst engineering students who routinely planned their course schedules around taking his classes. Dr. Agrawal is not only a Stanford trained head and neck surgeon, but an honorary engineer. His ability to quickly learn and apply technical concepts was astounding, and he consistently demonstrates a curiosity for knowledge in a wide variety of fields.

I would also like to thank the past and current members of the lab. Luke Helpard, you are the most positive and uplifting person I've ever met. Doing work alongside you provided me with a tremendous amount of energy. Alireza Rohani, not only do you have an impressive breadth of knowledge, but you consistently went out of your way to use that knowledge to help anyone who needed it. Hasitha Wimalarathna, Evan Simpson, Soodeh Nikan, Brad Gare, and Daniel Allen are all lab members who I deeply respect for unique reasons. I feel that I've grown tremendously both intellectually and personally as a result of my time with the lab, and I am extremely lucky to have shared these years with each of you.

Finally, I would like to thank my friends and family. Linda and John, you've supported me through every step of life. Your selflessness has afforded me endless opportunities which I am eternally grateful for. You are each my parent and my best friend, and I wouldn't have it any other way. Nicole, you open my mind to new ways of thinking and make me a better person each day. There is no one I'd rather figure out life with. Adam, Braedon, and Tiger, I am unable to do justice to what each of you mean to me in so few words. You've provided me with unwavering support, friendship, and motivation.

List of Tables

Table 3.1: Description of study data. ‘Sample Details’ lists data subsets which share image acquisition modalities. Each bullet point in ‘Input Acquisitions’ represents one acquisition and specifies if an electrode was implanted in the corresponding scans. Acquisitions used in the test set appear in bold. ‘Ground Truth Modality’ and ‘Ground Truth Segmentation Method’ describe the imaging modality and method used to form ground truth segmentations in the respective data subset. The number of samples assigned to train, validation, and test sets are specified in ‘Data Split’.........33

Table 3.2: Mean values are presented for all metrics. Data ranges are given in parenthesis. Metrics include Inner Ear Dice Coefficient (IE DC), Inner Ear Average Hausdorff Distance (IE AHD), Inner Ear Maximum Hausdorff Distance (IE MHD), and Round Window Centroid Distance (RW CD). Metrics are presented for three test set subsets as each is unique in clinical and ground truth modalities……...51

Table 4.1: Description of study data. ‘Sample Details’ lists data subsets which share image acquisition modalities. Each bullet point in ‘Input Acquisitions’ represents one acquisition and specifies if an electrode was implanted in the corresponding scans. Acquisitions used in the test set appear in bold. ‘Ground Truth Modality’ and ‘Ground Truth Segmentation Method’ describe the imaging modality and method used to form ground truth segmentations in the respective data subset. The number of samples assigned to train, validation, and test sets are specified in ‘Data Split’.........66

Table 4.2: Mean values are presented for all metrics. Data ranges are given in parenthesis. Metrics include Inner Ear Dice Coefficient (IE DC), Inner Ear Average Hausdorff Distance (IE AHD), and Inner Ear Maximum Hausdorff Distance (IE MHD). Metrics are presented for four test set subsets……72

Table 4.3: Quantitative results with std dev including Inner Ear Dice Coefficient (IE DC), Inner Ear Average Hausdorff Distance (IE AHD), Inner Ear Maximum Hausdorff Distance (IE MHD).........73

List of Figures

Figure 1.1: Cochlear implant (https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons)…………………………………………………………………………………......2

Figure 2.1: A subset of image pixels are shown up close without interpolation alongside their gray levels. The pixels shown on the left correspond to the pixels within the small red square on the right image..9

Figure 2.2: The left image depicts an RGB image of sprinters. The right image overlays a manual binary segmentation of image pixels classified as 'background' (blue) or 'person' (green)..........9

Figure 2.3: An affine transformation with translation and scaling steps is applied to image B to produce image A. A rigid transformation with translation and rotation steps is applied to image B to produce image C…………………………………………………………………………11

Figure 2.4: Coronal view of the temporal bone and surrounding structures. Image courtesy of Stanford Otolaryngology, Head and Neck Surgery. Illustration reprinted with permission: ©Chris Gralapp [3]..13

Figure 2.5: Electrode insertion into the cochlea. The implant follows the lateral wall along the inside of the cochlea. A similar approach is implemented whether insertion takes place in the RW or a cochleostomy. Image courtesy of Stanford Otolaryngology, Head and Neck Surgery. Illustration reprinted with permission: ©Chris Gralapp [3]..15

Figure 2.6: Surgical view of the IE through a mastoidectomy which exposes the RW. Image courtesy of Stanford Otolaryngology, Head and Neck Surgery. Illustration reprinted with permission: ©Chris Gralapp [3]...16

Figure 2.7: NN illustration………………………………………………………………………18

Figure 2.8: Convolution with 2D input and kernel……………………….…………………….……….22

Figure 2.9: The top row depicts images convolved with the kernels shown under each image. The leftmost image is the original; when convolved with the identity kernel, the image is unmodified. The middle image was convolved with a 3×3 box blur kernel, and the right image was convolved with a 3×3 sharpening filter…………………………………………………………….…….23

Figure 2.10: Demonstration of max pooling…………………………………………………….24

Figure 3.1: Axial slices of varying study data **a:** SR-PCI 9um isotropic, **b:** Micro-CT 40um isotropic, **c:** Helical CT 234×234×625um35…………………………………………………35

Figure 3.2: 2D visualization of the SR-PCI atlas-based rigid registration. **a:** The fixed image **(**helical CT with 234×234×625um voxel spacing)**, b:** The moving image (SR-PCI 9um isotropic voxel spacing) overlaid upon the fixed image after registration was applied. The resultant RW is displayed in green and identified by an arrow………………………………………………37

Figure 3.3: 3D visualization of the SR-PCI ground truth RW segmentation in cyan, and the atlas approach RW in fuchsia. **a,c:** atlas fits to target anatomy **b:** atlas differs from target anatomy. 3D visualization was done with 3DSlicer (v4.11, https://www.slicer.org/) [13].........................38

Figure 3.4: Expert segmentation post-processing. The left column depicts unprocessed expert segmentations, and the right column depicts post-processed segmentations. From top to bottom, segmentations correspond to: 300um isotropic voxel spacing CBCT, 234×234×625um helical CT, implanted 300um isotropic voxel spacing CBCT…………………………………………..40

Figure 3.5: Dice scores with and without post-processing……………………………………...40

Figure 3.6: Dice scores of processed expert segmentations and STAPLE consensus………….41

Figure 3.7: Axial view of 300um isotropic cadaveric CBCT scans displayed without interpolation. **a:** scan without artificial noise, **b:** scan with low artificial noise (standard deviation 185 HU), **c:** scan with high artificial noise (standard deviation 295 HU)....................................42

Figure 3.8: Axial view of voxel intensities sampled on a non-cadaveric CBCT scan without interpolation. **a:** a line through the temporal bone near the cochlea, **b:** a line through the fluid filled internal auditory canal……………………………………………………………………..43

Figure 3.9: 3D renderings of pipeline outputs for two test set non-cadaveric scans. The left column shows inferences from the model trained without noise augmentations. The right column shows inferences from the model with the noise augmentation……………………………….44

Figure 3.10: CNN architecture for 50um IE and RW segmentation……………………………..46

Figure 3.11: Average image intensities for 33 study scans including 17 non-cadaveric CBCT scans, eight cadaveric CBCT scans, and eight cadaveric helical CT scans……………………...48

Figure 3.12: 3D renderings of intermediate pipeline segmentations for a non-cadaveric CT. **a:** raw network output. Subtle patch lines are present as a result of the patch-based sliding window inference approach, **b:** the segmentation after Taubin smoothing. Small false positive islands are visible, **c:** the final pipeline output after island removal……………………………………...49

Figure 3.13: 2D and 3D visualizations of pipeline output segmentations. IE (depicted in green) and RW (depicted in yellow) can be seen in all eight test set scans. **a,b,c:** 234×234×625um, **d,e,f:** 300um isotropic, **g,h:** 300um isotropic electrode-implanted……………………………...52

Figure 3.14: Distances between the pipeline output and ground truth segmentations are depicted by coloured pipeline output segmentations for eight test set scans. Distances are in units of mm and follow the scale shown in image **a**. **a,b,c:** 234×234×625um, **d,e,f:** 300um isotropic, **g,h:** 300um isotropic electrode-implanted. 3D renderings and distance calculations were completed with MeshLab [24] (v 1.3.2, https://www.meshlab.net/)..53

Figure 3.15: Dice scores comparing processed expert segmentations, STAPLE consensus, and the automated pipeline……………………………………………………………………...54

Figure 3.16 Dice scores comparing processed expert segmentations to pipeline segmentations based on each test set data subset……………………………………………………….55

Figure 3.17: 3D renderings of ground truth segmentations, pipeline predictions, and expert segmentations. Each row corresponds to one of the eight cadaveric samples. Rows one to three correspond to 234×234×625um helical CT, rows four to six correspond to 300um isotropic CBCT, and rows seven and eight are 300um isotropic CBCT with implanted electrodes. The first column shows ground truth segmentations, the second column shows automated pipeline outputs, and columns three to six are expert segmentations……………………………………..56

Figure 3.18: 3D visualizations of pipeline output segmentations for all eight non-cadaveric test set scans. The IE is depicted in green, and RW in yellow……………………………………….58

Figure 3.19: 2D visualization of the average RW segmentation centroid error…………………59

Figure 3.20: 3D visualization of a pipeline output RW segmentation (in translucent yellow) with the predicted centroid in cyan, and the expert labeled centroid in fuchsia………………………59

Figure 4.1: Cadaveric helical CT study sample cropped to a 100mm diameter sphere surrounding the IE. From top to bottom and left to right: axial view, 3D rendering of bone-density scan voxels, coronal view, sagittal view. 2D and 3D visualization was done with 3DSlicer (v4.11, https://www.slicer.org/)...69

Figure 4.2: 3D renderings of pipeline output segmentations for all implanted test set scans. Pairs of side-by-side images correspond to a single sample; the right-hand image of each pair corresponds to the randomly rotated sample……………………………………………………73

Figure 4.3: 3D renderings of predicted segmentations for non-implanted test set scans. Pairs of side-by-side segmentations correspond to a single sample; the right-hand segmentation of each pair corresponds to the randomly rotated sample……………………………………………..74

Figure 4.4: 3D renderings of pipeline segmentations with and without rotational augmentations. The left and right columns show pipeline outputs with and without rotational augmentations respectively. The top row is based on cadaveric 300um isotropic CBCT, and the second row non-cadaveric CBCT………………………………………………………………………………76

List of Abbreviations

2D Two-dimensional

3D Three-dimensional

CBCT Cone Beam Computed Tomography

CI Cochlear implant

CNN Convolutional Neural Network

CT Computed Tomography

HU Hounsfield Units

IE Inner Ear

ML Machine Learning

Neural Network NN

PET Positron Emission Tomography

ReLU Rectified Linear Unit

RGB Red Green Blue

RW Round Window

SCC Semicircular Canals

SR-PCI Synchrotron Radiation Phase Contrast Imaging

STAPLE Simultaneous Truth and Performance Level Estimation

UHR-CT Ultra High-Resolution Computed Tomography

List of Appendices

Appendix A: Cadaveric Ethics Approval..81

Appendix B: Clinical Ethics Approval..82

Table of Contents

Abstract ii

Summary for Lay Audience iii

Co-Authorship Statement iv

Acknowledgements v

List of Tables vi

List of Figures vii

List of Abbreviations x

List of Appendices xi

Chapter 1 The Clinical Problem 1
- 1.1 Motivation 1
  - 1.1.1 Surgical Decision-Making 2
  - 1.1.2 Surgical Training and Rehearsal 3
  - 1.1.3 Robot-Assisted Surgery 4
  - 1.1.4 Customized CI Frequency Mapping 4
- 1.2 Novel Contributions 4
- 1.3 Thesis Outline and Objectives 5
- References 5

Chapter 2 Background 8
- 2.1 Digital Images 8
  - 2.1.1 Introduction to Digital Images 8
  - 2.1.2 Image Segmentation 9
  - 2.1.3 Registration, Interpolation, and Resampling 11
- 2.2 Clinical Background 12
  - 2.2.1 Anatomy 12
  - 2.2.2 Hearing Loss and Cochlear Implants 14
  - 2.2.3 Medical Imaging 16
- 2.3 Machine Learning 17
  - 2.3.1 Neural Networks 17

Training a Neural Network 19

Data Normalization and Regularization 20

2.3.2 Image Convolution 22

2.3.3 Convolutional Neural Networks 23

CNN Building Blocks 23

Patch-Based CNN 25

2.3.4 Transfer Learning 26

References 27

Chapter 3 A Deep Learning Solution for Automated Segmentation of the Inner Ear and Round Window on Electrode-Implanted and Non-Implanted Computed Tomography Scans 31

3.1 Introduction 31

3.2 Materials and methods 32

3.2.1 Image Acquisition 32

3.2.2 Data Preparation and Augmentations 36

Image Segmentation 36

Image Augmentation 41

3.2.3 Segmentation Pipeline 44

3.2.4 Evaluation Metrics 49

3.3 Results 50

3.3.1 Testing on Cadaveric Scans With and Without Implanted Electrodes 50

Comparison to Ground Truth Segmentations 50

Comparison to Expert Segmentations 53

3.3.2 Testing on Non-cadaveric Scans Without Electrodes 57

3.4 Discussion 60

References 61

Chapter 4 A Deep Learning Solution for Automated Segmentation of the Inner Ear on Electrode-Implanted and Non-Implanted Computed Tomography with a Large Field of View and Varying Orientation 65

4.1 Introduction 65

4.2 Materials and Methods 66

4.2.1 Image Datasets, Preparation, and Augmentation 66

4.2.2 Segmentation Pipeline 69

4.3 Results 71

4.3.1 Testing on Cadaveric Scans with Electrodes 71

4.3.2 Testing on Non-cadaveric Scans Without Electrodes 73

4.4 Discussion 75

References 77

Chapter 5 Conclusion 77

5.1. Summary 78

5.2. Future Work 79

References 80

Appendices 81

Curriculum Vitae 83

Chapter 1

The Clinical Problem

1.1 Motivation

Computed Tomography (CT) imaging is a prerequisite for most surgical procedures involving the head and neck. Otolaryngologists analyze CTs to determine the next steps in patient care. CTs are also a key tool otolaryngologists use in surgical planning to familiarize themselves with the patient’s specific anatomy and identify any abnormalities. To motivate this thesis, relevant anatomy and cochlear implants (CIs) are briefly described here, and described in further detail in Chapter 2. CIs are electronic devices which provide the sensation of sound to individuals with severe hearing loss. Within a CI, electrical impulses are generated to activate specific regions of an electrode array which is surgically inserted into the cochlea, the spiral shaped end organ of hearing (Figure 1.1). The round window is a membranous entry point into the cochlea, and is often the place of insertion when surgeons implant the electrode array into the cochlea.

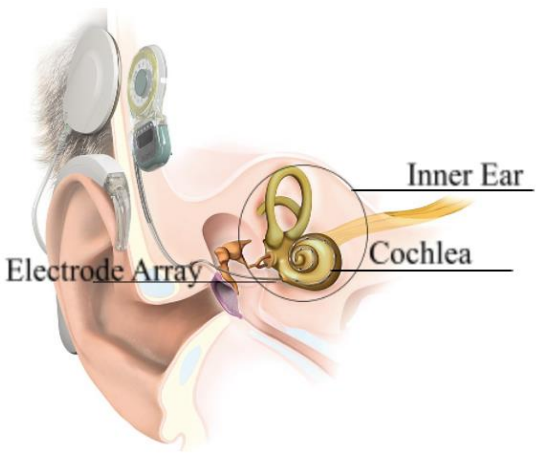


Figure 1.1: Cochlear implant (https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons).

Segmentation is the act of manually or automatically making pixel-wise classifications in an image. Segmentation of the inner ear (IE) and round window (RW) in clinical CT scans are useful for an ever-increasing number of applications which benefit CI recipients impacted by sensorineural hearing loss. Quick and accurate IE and RW segmentations warrant improvements to the CI procedure and CI sound perception quality. Manual segmentation of the IE and RW based on clinical CT requires extensive domain knowledge and is time consuming, making manual segmentation infeasible to integrate to most clinical practice. A system achieving quick and accurate segmentations of the IE and RW would be useful in the following applications which will be described in further detail:

- Surgical decision-making would be enhanced,
- Surgical training and rehearsal could be patient-specific,
- Robot-assisted surgery could be automated, and
- Customized cochlear implant frequency-mapping could be automated.

1.1.1 Surgical Decision-Making

Preoperative planning for the CI procedure currently involves a surgeon manually scrolling through two-dimensional (2D) image slices of a CT scan to build intuition as to the three-dimensional (3D) characteristics of the anatomy. When the IE can be interactively viewed as a 3D segmentation volume, anatomical understanding is improved, especially in cases of malformation [1], [2]. Improved anatomical understanding allows a surgeon to make more informed decisions. An optimal electrode length and insertion angle can be calculated based on the relative segmentations of the IE and RW [3]. Appropriate choices of insertion angle and electrode length optimize hearing outcomes and reduce residual hearing loss which can occur as a result of the implant damaging cochlear microstructures [4]–[6]. These findings incentivize an enhanced understanding of individual patient anatomy to achieve optimal CI insertion. In addition to preoperative scans, postoperative scans can also be useful in a clinical setting. If a postoperative complication arises such as complete or partial electrode extrusion from the IE, corrective surgery is often performed, in which case it is useful to visualize 3D segmentations of the IE and RW prior to surgical intervention [7]. CT scans with implanted electrodes pose additional challenges due to metallic artifacts which distort the anatomy surrounding the titanium implant casing [8], [9]. Metallic artifacts hinder visualization of much of the cochlea and RW, especially in cases where RW insertion was performed. A system which could accurately segment the RW in electrode-implanted scans would be especially useful due to the increased difficulty of interpreting electrode-implanted scans.

1.1.2 Surgical Training and Rehearsal

Surgical training and rehearsal methods across medical disciplines have increasingly utilized computer-based simulated training systems; the systems can involve any combination of virtual reality headsets, 3D monitors, and haptic feedback devices [10]. Virtual training environments have gained recent attention as a possible approach to maintain surgical skills while many non-essential surgical procedures have been stopped during the COVID-19 pandemic [11]. A limitation of many surgical simulation platforms is the inability to quickly import patient specific models corresponding to a medical image. CardinalSim is a surgical simulation platform being developed in the Auditory Biophysics Lab at Western University for temporal bone drilling, which is a necessary step in CI surgery [12]. Automated segmentations of the IE and RW facilitate quick implementation of patient cases to a simulated environment such as CardinalSim, allowing the procedure to be practiced by surgeons prior to the procedure with no risk. Complex cases could also be presented as a training example for surgical residents gaining proficiency in the procedure.

1.1.3 Robot-Assisted Surgery

The use of robotic systems is being explored for a variety of precision surgical tasks, including head and neck surgery; the performance of these systems has in some cases exceeded that of experts [13]. If a robotic system is to be applied to the CI procedure, patient specific segmentations would be necessary to give the device sufficient information to deduce safe drilling locations which avoid critical structures. Similarly to the manual CI procedure, the relative positions of the cochlea and RW could be used to determine an ideal electrode insertion angle for a robotic system.

1.1.4 Customized CI Frequency Mapping

Current approaches to stimulate CIs are one-size-fits all and take no consideration to the specific anatomy of the patient. Customized CI frequency-maps which account for patient anatomy have recently shown improved hearing outcomes for CI recipients, especially in cases where patient anatomy deviates significantly from the average [14], [15]. The manual steps involved in creating a customized frequency-map are time consuming and could be greatly reduced through automated segmentations. Automating manual segmentation, a time-consuming step in customized CI frequency-mapping, could help facilitate integration of the technique to routine clinical practice.

1.2 Novel Contributions

This is the first known study which achieves automated segmentation of the RW in electrode-implanted temporal bone CT scans, and the first study which achieves automated segmentation of the IE in electrode-implanted temporal bone CT scans. The effectiveness of the approach is evaluated by comparison with manual neurotologist and neuroradiologist segmentations. Segmentation accuracy metrics for IE and RW segmentation in non-implanted samples indicate competitive or improved results compared to other work. This work demonstrates superior robustness to other studies as it is validated by extensive CT data including clinical data from many

CT scanners as well as cadaveric micro-CT, Cone Beam Computed Tomography (CBCT), and helical CT. This work is also the first to validate a RW segmentation approach with synchrotron radiation phase contrast imaging (SR-PCI) data, and the first known study which segments the RW using a convolutional neural network (CNN).

1.3 Thesis Outline and Objectives

This thesis is structured as follows: Chapter 2 gives background information on concepts which are foundational to the thesis. Chapter 3 presents a method for automated high-resolution segmentation of the IE and RW in electrode-implanted and non-implanted cochleae. A limitation of this work was that the model was trained on data manually cropped to a small region of interest, and in a standardized clinical orientation. Chapter 4 presents a solution to the limitation, wherein a method is presented to segment the IE at a lower resolution in large field of view scans of any directional orientation with and without implanted electrodes. Together, Chapter 3 and Chapter 4 lay the groundwork for a fully automated IE and RW segmentation in a variety of scans. The main objective of this thesis was to develop, implement, and evaluate an automated segmentation algorithm for the IE and RW.

References

[1] S. Alenzi, A. Dhanasingh, H. Alanazi, A. Alsanosi, and A. Hagr, "Diagnostic Value of 3D Segmentation in Understanding the Anatomy of Human Inner Ear Including Malformation Types," *Ear. Nose. Throat J.*, vol. 100, no. 5_suppl, pp. 675S-683S, Sep. 2021, doi: 10.1177/0145561320906621.

[2] K. A. Powell, T. Liang, B. Hittle, D. Stredney, T. Kerwin, and G. J. Wiet, "Atlas-Based Segmentation of Temporal Bone Anatomy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 11, pp. 1937–1944, Nov. 2017, doi: 10.1007/s11548-017-1658-6.

[3] H. A. Breinbauer and M. Praetorius, "Variability of an ideal insertion vector for cochlear implantation," *Otol. Neurotol. Off. Publ. Am. Otol. Soc. Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol.*, vol. 36, no. 4, pp. 610–617, Apr. 2015, doi: 10.1097/MAO.0000000000000719.

[4] A. Khater and M. W. El-Anwar, "Methods of Hearing Preservation during Cochlear Implantation," *Int. Arch. Otorhinolaryngol.*, vol. 21, no. 3, pp. 297–301, Jul. 2017, doi:

10.1055/s-0036-1585094.

[5] H. L. Cornwall, P. S. Marway, and M. Bance, “A Micro-Computed Tomography Study of Round Window Anatomy and Implications for Atraumatic Cochlear Implant Insertion,” *Otol. Neurotol. Off. Publ. Am. Otol. Soc. Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol.*, vol. 42, no. 2, pp. 327–334, Feb. 2021, doi: 10.1097/MAO.0000000000002924.

[6] Y. Shapira, A. A. Eshraghi, and T. J. Balkany, “The perceived angle of the round window affects electrode insertion trauma in round window insertion - an anatomical study,” *Acta Otolaryngol. (Stockh.)*, vol. 131, no. 3, pp. 284–289, Mar. 2011, doi: 10.3109/00016489.2010.533698.

[7] N. Vaid, J. T. Roland, and S. Vaid, “Extracochlear electrode extrusion,” *Cochlear Implants Int.*, vol. 12, no. 3, pp. 177–180, Aug. 2011, doi: 10.1179/146701010X12711475887234.

[8] H. Demirturk Kocasarac *et al.*, “Evaluation of artifacts generated by titanium, zirconium, and titanium-zirconium alloy dental implants on MRI, CT, and CBCT images: A phantom study,” *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.*, vol. 127, no. 6, pp. 535–544, Jun. 2019, doi: 10.1016/j.oooo.2019.01.074.

[9] T. Stöver and T. Lenarz, “Biomaterials in cochlear implants,” *GMS Curr. Top. Otorhinolaryngol. Head Neck Surg.*, vol. 8, p. Doc10, Mar. 2011, doi: 10.3205/cto000062.

[10] S. S. Y. Tan and S. K. Sarker, “Simulation in surgery: a review,” *Scott. Med. J.*, vol. 56, no. 2, pp. 104–109, May 2011, doi: 10.1258/smj.2011.011098.

[11] T. Doulias, G. Gallo, I. Rubio-Perez, S. O. Breukink, and D. Hahnloser, “Doing More with Less: Surgical Training in the COVID-19 Era,” *J. Invest. Surg.*, vol. 35, no. 1, pp. 171–179, Jan. 2022, doi: 10.1080/08941939.2020.1824250.

[12] E. C. Compton *et al.*, “Assessment of a virtual reality temporal bone surgical simulator: a national face and content validity study,” *J. Otolaryngol. - Head Neck Surg.*, vol. 49, no. 1, p. 17, Dec. 2020, doi: 10.1186/s40463-020-00411-y.

[13] A. B. Auinger, V. Dahm, R. Liepins, D. Riss, W.-D. Baumgartner, and C. Arnoldner, “Robotic Cochlear Implant Surgery: Imaging-Based Evaluation of Feasibility in Clinical Routine,” *Front. Surg.*, vol. 8, 2021, Accessed: Mar. 10, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fsurg.2021.742219

[14] L. Helpard *et al.*, “An Approach for Individualized Cochlear Frequency Mapping Determined From 3D Synchrotron Radiation Phase-Contrast Imaging,” *IEEE Trans. Biomed. Eng.*, vol. 68, no. 12, pp. 3602–3611, Dec. 2021, doi:

10.1109/TBME.2021.3080116.

[15] N. T. Jiam, M. Gilbert, J. Mo, P. Jiradejvong, and C. J. Limb, “Computed Tomography–Based Measurements of the Cochlear Duct: Implications for Cochlear Implant Pitch Tuning,” *Ear Hear.*, vol. 42, no. 3, pp. 732–743, Jun. 2021, doi: 10.1097/AUD.0000000000000977.

Chapter 2

Background

2.1 Digital Images

2.1.1 Introduction to Digital Images

Digital images are composed of a sequence of finite, discrete values known as pixels in 2D, or voxels (volumetric pixels) in 3D. Pixels take on a range of values corresponding to the bit depth of the image with a relationship of: $range = 2^b$, where b represents the bit depth. In a grayscale image, the intensity of a given image pixel is known as it's gray level. For an 8-bit image, gray levels take on a range of $2^8 = 256$ possible values (0 to 255 for unsigned integer representations). Figure 2.1 shows an 8-bit grayscale image which demonstrates how digital images appear continuous but are discrete image pixels when observed closely (if visual interpolation techniques are not used). The dimensionality of a digital image depends on whether an image is grayscale or 'red green blue' (RGB). 2D grayscale images have a single channel dimension along with their associated spatial dimensions of length and width (1×L×W) whereas 2D RGB images have 3 channels corresponding to red, green, and blue, in addition to spatial dimensions (3×L×W).

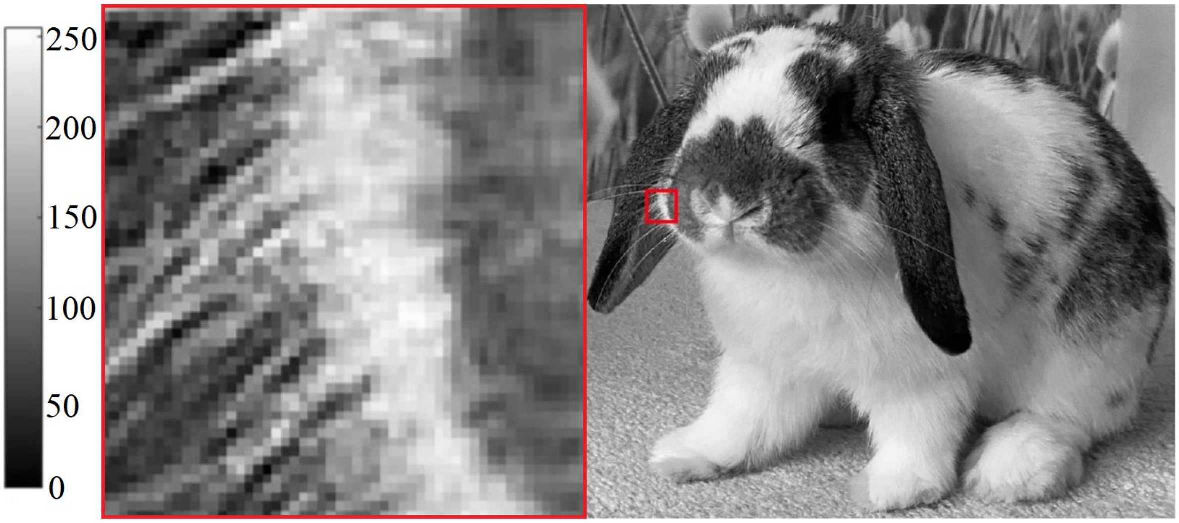


Figure 2.1: A subset of image pixels are shown up close without interpolation alongside their gray levels. The pixels shown on the left correspond to the pixels within the small red square on the right image.

2.1.2 Image Segmentation

Image segmentation involves the partitioning of an image to two or more groups, which is achieved through pixel-wise classification of the image. Figure 2.2 depicts an example image and the corresponding binary segmentation where each pixel is classified to label 0 (background) or 1 (person).

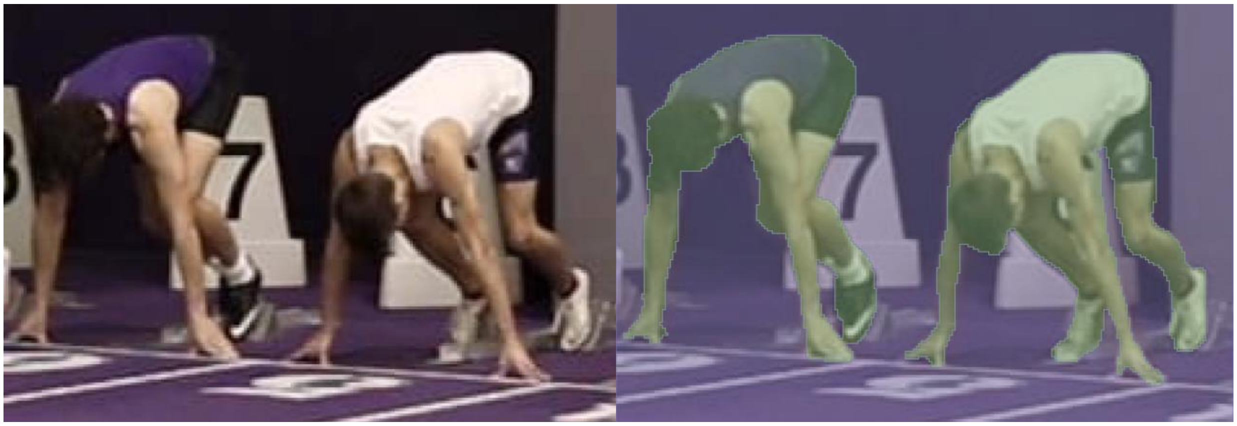


Figure 2.2: The left image depicts an RGB image of sprinters. The right image overlays a manual binary segmentation of image pixels classified as ‘background’ (blue) or ‘person’ (green).

Image segmentations have an identical spatial dimensionality to the images upon which they are based. Segmentation pixels take on values which correspond to the number of classes segmented. In the binary case, where only background and foreground classes are defined, the possible values of a segmentation are limited to 0 or 1. To segment outdoor scenery images, someone may decide to classify each image pixel as one of four categories: 0 (ground), 1 (tree), 2 (sky), 3 (other).

Accurate segmentations of human anatomy in medical scans are useful for an increasingly large number of applications; for example, they can be used to quantify tumor shrinkage in a patient receiving cancer treatment. Segmentations of human anatomy can be produced based on medical image scans such as 2D planar X-ray, as well as 3D CT and MRI. Manual segmentations of human anatomy are time consuming and require expert levels of domain knowledge to correctly interpret medical image data.

Image segmentation can be done manually, semi-automatically, or automatically. Manual segmentation involves a user classifying pixels with segmentation software; many software tools have been developed for the segmentation of 3D images [1]. Manual segmentation is the most time-consuming approach, as the user must carefully select pixels corresponding to structures of interest. Semi-automated segmentation approaches involve user input as well as an algorithm. In the semi-automated region growing segmentation approach, users select a seed pixel, and neighboring pixels are iteratively evaluated to check if they meet a similarity criteria. The segmenter can then add more seed pixels or manually correct errors made by the region growing algorithm. A benefit of manual or semi-automated segmentation is that human knowledge can guide decision making. For example, if the delineation between structures in a medical image is unclear due to an imaging artifact or poor contrast, an expert may be able to infer the location of structures by relying on their anatomical knowledge. Automated segmentation approaches involve segmentation with the least user input; the CNN is one such approach. Automated approaches sometimes require preprocessing steps to ensure the input is in a standardized format, as well as post-processing steps to finalize outputs. The pre- and post-processing steps may be done manually (making the overall approach semi-automated) or automated. Some important factors to consider when evaluating the effectiveness of a segmentation approach are:

- How long do automated and manual steps take?
- What level of knowledge is necessary to complete manual steps?
- What computational resources are required?

2.1.3 Registration, Interpolation, and Resampling

Image registration is a technique used to align images. Registration is achieved by bringing multiple images into one coordinate system. The image which has its coordinate system altered is referred to as the moving image, and the image which has the target coordinate system is the fixed or target image. The transformation applied to the moving image is determined by the type of registration which is applied. Rigid registration limits the transformation to only translation and rotation; affine registration additionally allows for scaling and shearing transforms. Figure 2.3 demonstrates rigid and affine transformations. The restrictions associated with each transformation method form the search space of possible transformations to achieve registration [2]. Affine and rigid transformations are applied to the entirety of images; elastic transformations, on the other hand, allow for local warping of images.

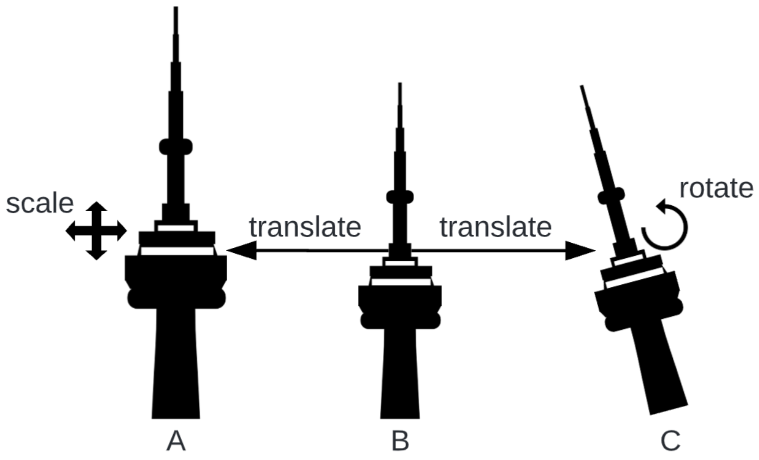


Figure 2.3: An affine transformation with translation and scaling steps is applied to image B to produce image A. A rigid transformation with translation and rotation steps is applied to image B to produce image C.

Registration techniques can also be classified as intensity-based or feature-based. Intensity-based techniques achieve registration through comparison of image pixel values in the moving and target images. Feature-based techniques consider image features such as lines or points which may be determined manually or automatically. If the correspondence between moving and target image features are known, the registration is achieved by finding a transformation which minimizes the distance between corresponding features in the moving and target images. To register a set of corresponding points on moving and target images, a cost function (e.g., the squared Euclidean distance between corresponding points) is minimized. The transformation corresponding to this minimization is the transformation matrix applied to the moving image.

Following a transformation, interpolation is necessary to bring the transformed image into the image grid of the fixed image. To assign values to fixed image grid pixels, nearby pixels on the transformed image are considered. Nearest neighbor interpolation is the simplest as it only considers the nearest transformed image pixel to the target image grid pixel. Linear and cubic interpolation approaches consider the influence of multiple pixels in the transformed image and lead to qualitatively preferential results in most cases.

Resampling an image involves changing the pixel size and number of pixels in an image. Upsampling refers to decreasing the size of pixels and therefore increasing the number of pixels, and downsampling involves increasing the size of pixels and decreasing the total number of pixels in an image. Interpolation techniques can be used to determine new pixel values in an upsampled or downsampled image.

2.2 Clinical Background

2.2.1 Anatomy

The temporal bones are paired bones which form a lateral inferior portion of the skull that protects the temporal lobe of the brain and houses organs responsible for hearing and balance. The external auditory canal (ear canal), middle ear (malleus, incus, stapes), and IE (cochlea and semicircular canals) are located within the temporal bone (Figure 2.4). The carotid artery and jugular veins also pass through the temporal bone, which facilitate most blood flow to and from the brain.

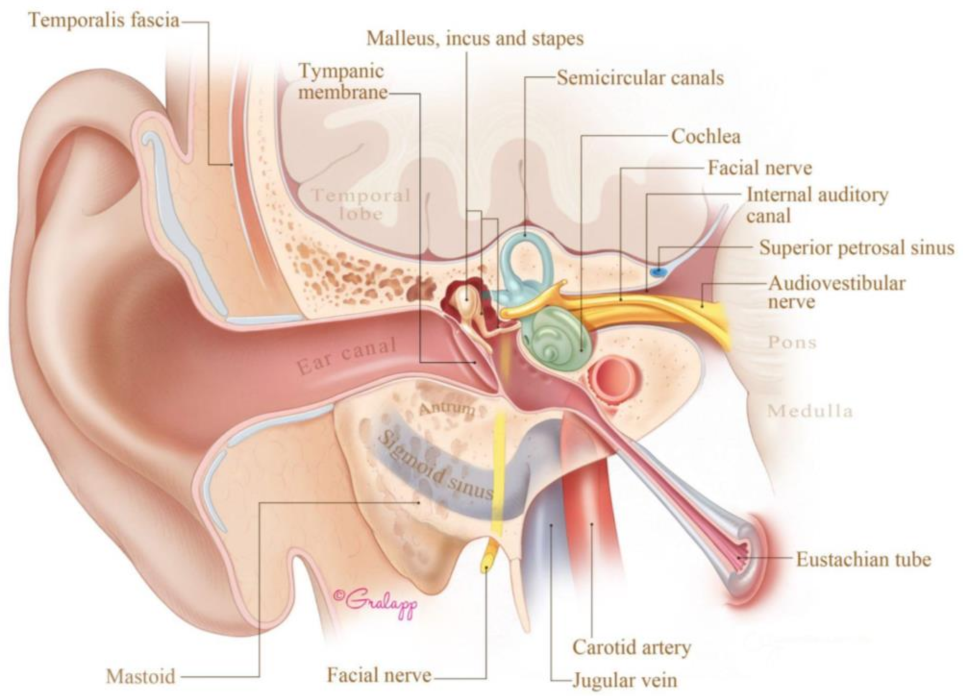


Figure 2.4: Coronal view of the temporal bone and surrounding structures. Image courtesy of Stanford Otolaryngology, Head and Neck Surgery. Illustration reprinted with permission: ©Chris Gralapp [3].

The IE is a bony labyrinth located in the temporal bone. The IE comprises the cochlea, vestibule, and three semicircular canals (SCC) (superior, horizontal, posterior). The anterior portion of the IE is the cochlea, a spiral shaped cavity which is responsible for the conversion of sound waves to nerve impulses that are interpreted by the brain. The cochlea averages 2.5 turns around the conical bony modiolus. The modiolus houses the cochlear nerve’s neuronal cell bodies which carry electrical signals to the brain. The cochlea has three internal channels including the most inferior scala tympani, the middle scala media, and the superior scala vestibuli (Figure 2.6). The basilar membrane separates the scala tympani from the scala media and vibrates in response to acoustic stimuli. The vestibule is a bony cavity posterior to the cochlea and contains the utricle and saccule which help with balance. The three semicircular canals are posterior to the vestibule and also contribute to balance by sensing perilymph fluid motion. The RW is the most anterior of two openings between the inner and middle ear, located at the basal end of the cochlea. The RW allows for perilymph motion in the cochlea by vibrating with opposite phase to vibrations in the IE. The RW averages just 1.8mm in diameter and 70um in thickness [4].

2.2.2 Hearing Loss and Cochlear Implants

Worldwide 432 million adults and 34 million children require rehabilitation to address their disabling hearing loss [5]. Disabling hearing loss is quantified as a loss of greater than 35 decibels in the better hearing ear. The negative impacts of hearing loss on one's life are extensive if left untreated; those impacted disproportionately face social isolation as well as poor educational and employment outcomes [6].

Hearing loss can be categorized as sensorineural, conductive, or mixed. Sensorineural hearing loss is the most common form of hearing loss and involves an issue with the IE or auditory nerve. Sensorineural hearing loss is commonly caused by aging or noise damage. Conductive hearing loss involves an issue with sound conduction in the outer or middle ear structures such as the ear canal, ear drum, and ossicles (malleus, incus, stapes). Common causes of conductive hearing loss include cerumen (earwax) buildup in the external auditory canal, perforation of the tympanic membrane (eardrum), or infection. Mixed hearing loss describes conditions which could be related to a combination of the inner, middle, and outer ear. An example of mixed hearing loss is an infection which is present in both the middle and IE.

A cochlear implant is an electronic device composed of an electrode array which, when surgically inserted into the cochlea, can be used to provide the sensation of sound to individuals with sensorineural hearing loss by directly stimulating cochlear nerve fibers [7]. CIs may be considered in cases of severe sensorineural hearing loss where a patient struggles with conversation even with the use of a hearing aid and relies heavily on lip reading. CIs may be implanted bilaterally (in both cochleae), or unilaterally (in one side). Bilateral implantation results in better hearing outcomes and allows for improved sound localization, as the difference in time and intensity of perceived sounds between the ears allows for understanding of the direction of a sound source [8], [9]. A CI is inserted into the cochlea through the RW or a cochleostomy, an unnatural entrance to the cochlea achieved by drilling (Figure 2.5). To access the cochlea, a mastoidectomy is performed, in which a portion of the mastoid part of the temporal bone is drilled away to expose the RW or a site for cochleostomy. The CI is inserted into the scala tympani, the inferior-most duct in the cochlea (Figure 2.6). Insertion is performed carefully to avoid puncture of the basilar membrane, which separates the scala tympani from scala media, as damage could result in residual hearing loss. Residual hearing is the natural hearing a person still has when hearing loss is present. Preventing residual hearing loss during CI surgery is important, as demonstrated by the development of techniques to monitor changes to residual hearing intraoperatively with electrocochleography [10], [11].

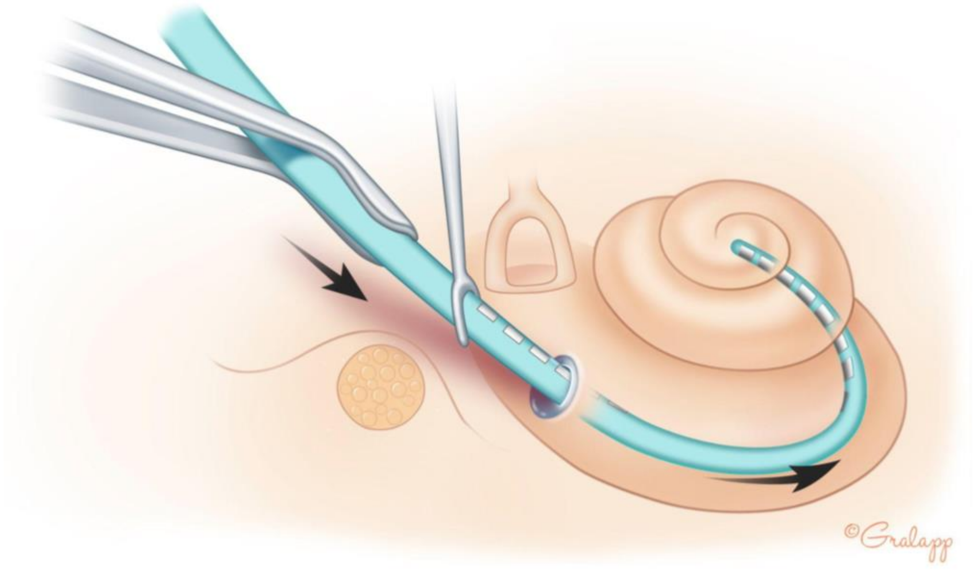


Figure 2.5: Electrode insertion into the cochlea. The implant follows the lateral wall along the inside of the cochlea. A similar approach is implemented whether insertion takes place in the RW or a cochleostomy. Image courtesy of Stanford Otolaryngology, Head and Neck Surgery. Illustration reprinted with permission: ©Chris Gralapp [3].

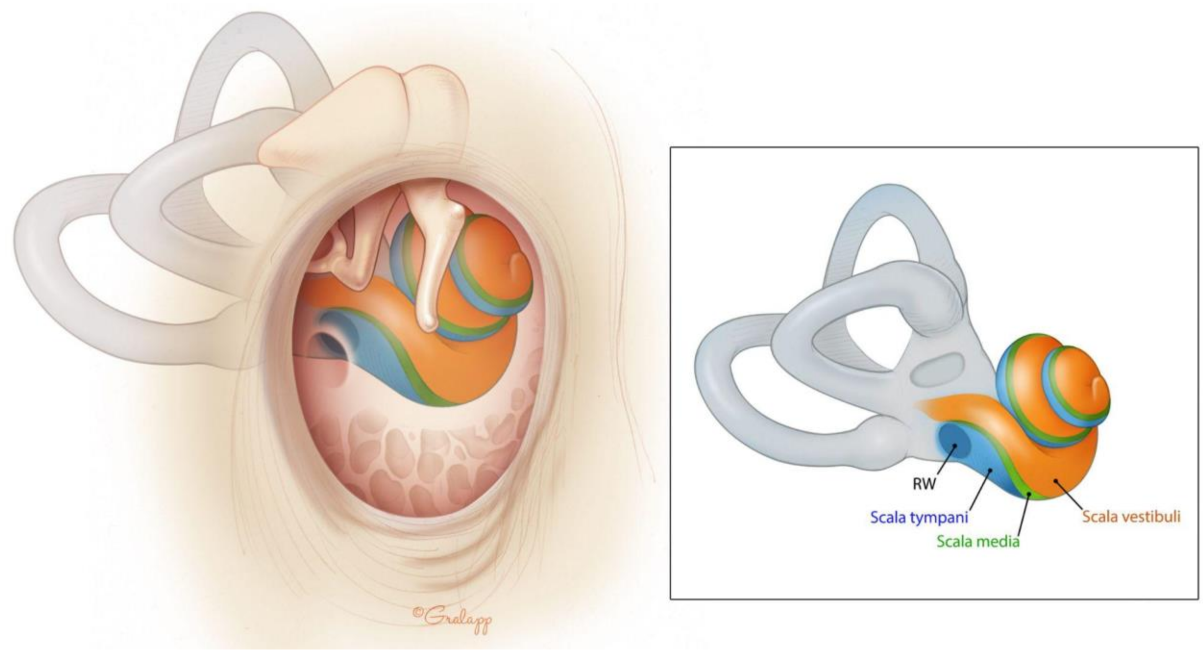


Figure 2.6: Surgical view of the IE through a mastoidectomy which exposes the RW. Image courtesy of Stanford Otolaryngology, Head and Neck Surgery. Illustration reprinted with permission: ©Chris Gralapp [3].

The facial nerve (seventh cranial nerve) may be located close to or in the ideal electrode insertion path depending on patient anatomy. Damage to the nerve could result in irreversible facial paralysis. A surgeon must be especially careful in cases of ossification or fibrosis of the IE, where growth of bone or soft tissue may obscure surgical landmarks in the basal turn of the cochlea. Excessive drilling superiorly could result in damaging the cochlear nerve in the modiolus, and anteriorly could damage the intratemporal carotid artery.

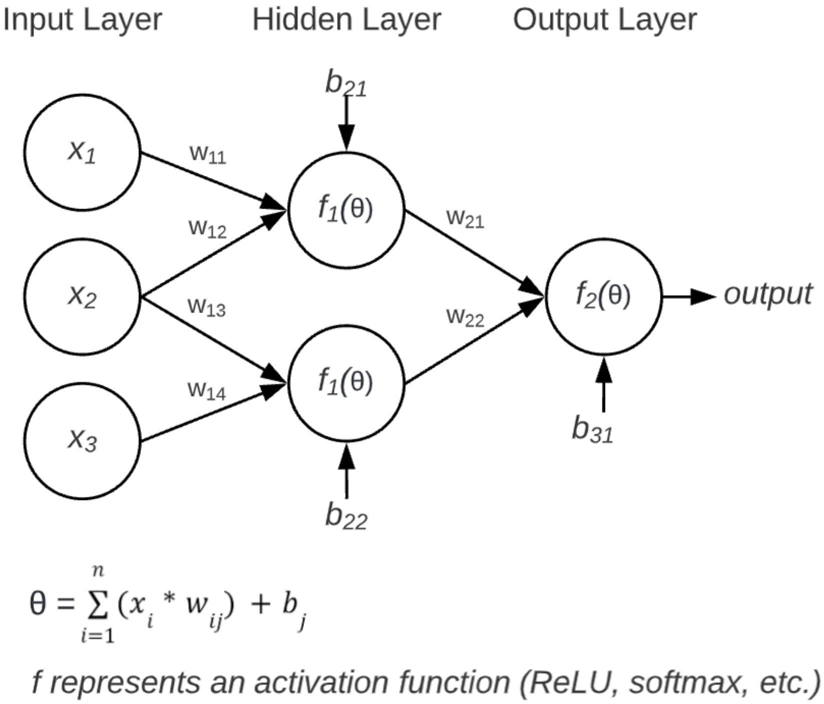
2.2.3 Medical Imaging

With the advent of medical imaging techniques came incredible improvements to healthcare. The ability to quickly and non-invasively view internals of the human body allowed for improvements in medical diagnostics, surgical planning, pathology, and anatomy. Medical images are 2D or 3D digital images which represent human anatomy or function. In CT and X-ray, image intensities correspond to the density of objects. In positron emission tomography (PET), intensities correspond to radiation emitted from radioisotopes introduced to the body. In magnetic resonance imaging, intensities correspond to electron spins induced by a magnetic field. Medical images of different modalities which pertain to a specific subject can be aligned with registration techniques; an especially useful application of this is the alignment of a PET scan to a CT scan which combines information to show detailed structures and cellular activity within the body [12]. The combination of CT and PET scans was found to be useful, leading to the development of PET-CT scanners which acquire both PET and CT images in one scanning session [13]. CT was the first 3D medical imaging modality and shares many characteristics with the planar X-ray. The first commercially viable CT scanner was invented by Godfrey Hounsfield in 1972; as a result, voxel values in a CT scan are referred to as Hounsfield units (HU) which are a linear transformation of the estimated linear attenuation coefficient. The linear attenuation coefficient corresponds to the estimated density of an area within the body. Because dense tissues attenuate the transmission of X-rays more than less dense tissue, visualizing the densities results in an image which discriminates structures within the body. By rotating an X-ray source and detector around a patient, as is done in CT, information can be gathered to determine the density of structures in 3D. Filtered back projection or iterative algorithms are image reconstruction approaches to convert raw CT sinogram data to useful image data. The quality of a CT scan is impacted by spatial resolution, contrast, noise, and distortion. Micro-CT is an imaging approach which uses the same fundamentals as CT, but achieves higher resolution, improved contrast, and reduced noise. Micro-CT scanners are not used on live patients as they expose scanned specimens to dangerous levels of radiation, and because they can generally only be used with small specimens, i.e., the use of micro-CT requires dissection.

2.3 Machine Learning

2.3.1 Neural Networks

Neural networks (NNs) are a subset of machine learning (ML) algorithms which involve neurons or nodes which learn underlying relationships in data. They are the focus of this section as the thesis utilizes them. The name originates from biological neurons which are electrically excitable cells in the brain. NNs involve layers of nodes connected to one another; layers are categorized as input layers, hidden layers, or output layers (Figure 2.7). NNs have a single input layer which acts as the entry point for all data fed into the network. Networks may have any number of hidden layers and nodes per layer. There is a single output layer which may produce a continuous value in the case of a regression problem like forecasting product sales, or a discrete value in the case of classifying objects in images. Each connection between nodes has an associated weight, and each node has a bias term. Nonlinear activation functions such as rectified linear unit (ReLU) or sigmoid are applied to neuron activations to allow models to represent nonlinearity [14]. The activation of any given node is calculated as the summation of all inputs times their respective weights plus the node's bias term, which is then input to an activation function ($node\ activation = f(\sum_{i=1}^{n}(x_i w_{ij}) + b_j\ )$) where x is the incoming activation, w is the weight associated with that activation, and b is the bias term of the node. The i and j subscripts indicate which node the activations, weights, and biases refer to. The activation function of the output layer depends on the type of ML problem, and often differs from the activation function used in hidden layers. For classification problems, the softmax and sigmoid functions are used to convert logits to probabilities which sum to 1; the probabilities correspond to the network's level of confidence that an input belongs to each class.



$$\theta = \sum_{i=1}^{n}(x_i * w_{ij}) + b_j$$

f represents an activation function (ReLU, softmax, etc.)

Figure 2.7: NN illustration.

Training a Neural Network

A loss function determines the amount of error associated with a network prediction. A high loss indicates that a prediction was poor, and a low loss indicates that the prediction was accurate. In the case of semantic segmentations, **Dice loss** can be used. Dice loss is calculated as one minus the Dice coefficient $(1-\frac{2|X\cap Y|}{(|X|+|Y|)})$ for segmentations X and Y, and measures the amount of overlap between a predicted and target segmentation. The numerator of the Dice coefficient is the number of overlapping voxels between the segmentations, and the denominator is the sum of each segmentation's voxels individually. When segmentations completely overlap, the Dice coefficient is 1, and Dice loss is 0. When segmentations have no overlap, the Dice coefficient is 0 and the Dice loss is 1. **Cross-entropy loss** $(-log(p_t))$ is used to compare probabilities (p_t) output by the network to known labels for the data. For example, in a classifier which labels images as cats or dogs, a model may have a confidence of 0.7 that an image of a dog is a dog, which corresponds to a cross-entropy loss of $-log(0.7) = 0.155$. If the model only had a confidence of 0.2 that an image of a dog was a dog (the model was 80% confident that it was a cat), the cross-entropy loss would be: $-log(0.3) = 0.699$. **Focal loss** $(-(1-p_t)^{\gamma} log(p_t), \gamma = 2)$ is an extension to cross-entropy loss which adds a term $(1-p_t)^{\gamma}$ to cross-entropy loss. Focal loss exacerbates the loss contribution of poorly classified examples [15]. γ is a regularization term which controls the contribution of the focal loss factor. With the same example, the focal losses are: $-(1-0.7)^2 log(0.7)) = 0.014$, and $-\left(1-0.3)^2 log(0.3)\right) = 0.256$. In the case of cross-entropy loss, the contribution of the poorly classified example is 4.5× greater than the well classified example, in the case of focal loss, the poorly classified example yields a loss 18.2× larger.

The training of NNs involves the minimization of a loss function; often this minimization is done using **gradient descent**, and it is achieved through the backpropagation algorithm. When an example is fed into a NN, the model makes a prediction, and a loss function estimates error by comparing the model prediction to a known ground truth. In **backpropagation**, network prediction errors are used to update weights and biases by estimating the contribution of each weight and bias to the overall error. The name backpropagation refers to the fact that weights and biases are updated starting from the output layer, and information is propagated backwards towards the input layer. Partial derivatives are used at each layer to estimate the contribution of each parameter to loss, and to calculate updates for earlier layers. The **batch size** of a NN defines the number of examples it considers before making updates to weights and biases. One **iteration** is defined as the completed forward pass and backwards pass (backpropagation) associated with a batch. An **epoch** is defined as the network seeing each example of the dataset once. For example, if there are 100 training examples and a batch size of two, one epoch involves 50 iterations. The **learning rate** of a NN is a scaling factor which determines how much weights and biases are updated with each iteration. A learning rate which is too low could lead to a model which converges slowly or is unable to escape local minima in loss, and a learning rate which is too high may lead to oscillatory non-convergence where minima in loss are missed.

To train ML models including NNs, data is split into training, validation, and test sets. The training set is used by the network to update parameters (weights and biases). The validation set is used to evaluate the accuracy of the model with data that did not contribute to the update of model parameters. The validation set may be used to estimate model accuracy following one or more epochs. Finally, the test set is used to report accuracy of a model. Test set data was not used to update model parameters, like the training set, or to select the ideal model, like the validation set.

Data Normalization and Regularization

Data is often normalized or standardized before being fed into a machine learning model. Normalization allows input features with differing ranges to contribute equally to learned model parameters. Normalization can be achieved by transforming all values to a range between zero and one, and standardization by transforming data to have zero mean and unit standard deviation. For example, in a model aiming to predict life expectancy with inputs of age (ranging from 0 to 100) and salary (ranging from 0 to 1,000,000), the age and salary inputs should be normalized or standardized due to the relative scale of values. Normalization may also occur within a NN in the form of batch or group normalization; these techniques add trainable parameters to a NN which normalizes node activations. Research as to why batch and group normalization are effective is still ongoing, but it is thought that the techniques smoothen the loss function which is traversed in gradient descent. In batch normalization, activations are normalized along the batch dimension

[16]. Group normalization was introduced as an alternative to batch normalization which normalizes independently of batches. Group normalization is often preferred for applications such as computer vision, which utilizes small batch sizes due to memory limitations [17].

Regularization techniques are methods which aim to avoid overfitting in NNs. Overfitting occurs when a model begins to memorize training data and loses generalizability. Overfit models are considered to have high variance, as their performance heavily depends upon which subset of data is used to test the model. Overfitting can be identified when training loss continues to drop, but validation loss is rising. Some common regularization techniques are early stopping, dropout, L1 and L2 regularization, and data augmentation; the previously described batch and group normalization also have regularization effects on a network [18]. In early stopping, the model which achieves the lowest validation loss is selected as the ideal model; this technique requires that the dataset is split into train and validation sets as was previously described. Dropout involves randomly temporarily removing a subset of network nodes during a given training iteration. Randomly removing nodes from the network prohibits the network from relying heavily on specific nodal pathways, which is frequently present in cases of overfitting. L1 and L2 regularization add a term to the loss function which punishes large weights. Functionally, punishing large weights prohibits specific nodes from dominating the network. L1 regularization adds a coefficient $(|w|)$ to the loss function which is the absolute value of network weights, and L2 regularization adds a coefficient (w^2) which is the square of the weights. Compared to L1 regularization, L2 regularization disproportionately punishes large weights, and does not punish small ones. L1 regularization accomplishes feature selection as small weights are forced to zero. Data augmentation reduces overfitting by artificially increasing the number of examples in a dataset. If an insufficient number of examples are provided to a NN, the network may learn a high variance function which minimizes loss by memorizing examples [19]. Applying augmentations such as translation, flip, scaling, and rotation to image data may result in a more robust model, especially when data is limited.

2.3.2 Image Convolution

Convolutional kernels are general purpose filters which achieve a variety of tasks depending on kernel values. Convolutional kernels are applied to images by calculating a weighted sum of image pixels, with the weights corresponding to kernel values. The stride associated with a convolution describes how much the convolutional kernel is shifted across the image between each convolutional step. With a stride larger than one, downsampling of the image is achieved. Padding refers to the addition of zeros or mirrored pixel values surrounding the border of an image prior to convolution, which allows the output following convolution to match the input image size. Figure 2.8 demonstrates a 3×3 convolutional kernel with a stride of one being applied to a 5×5 image with no padding.

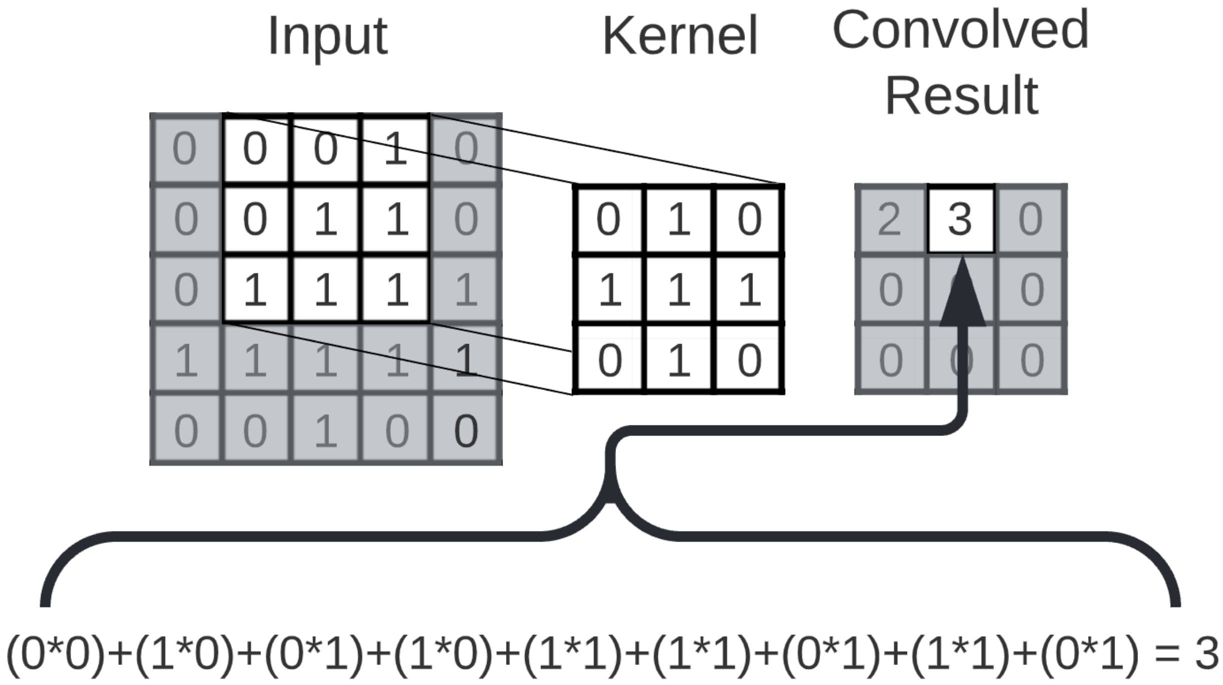


Figure 2.8: Convolution with 2D input and kernel.

As convolutional kernels are general purpose filters, they can achieve a wide variety of tasks such as edge detection, sharpening, and blurring. Figure 2.9 demonstrates the diverse impacts of applying convolutional kernels to an image.

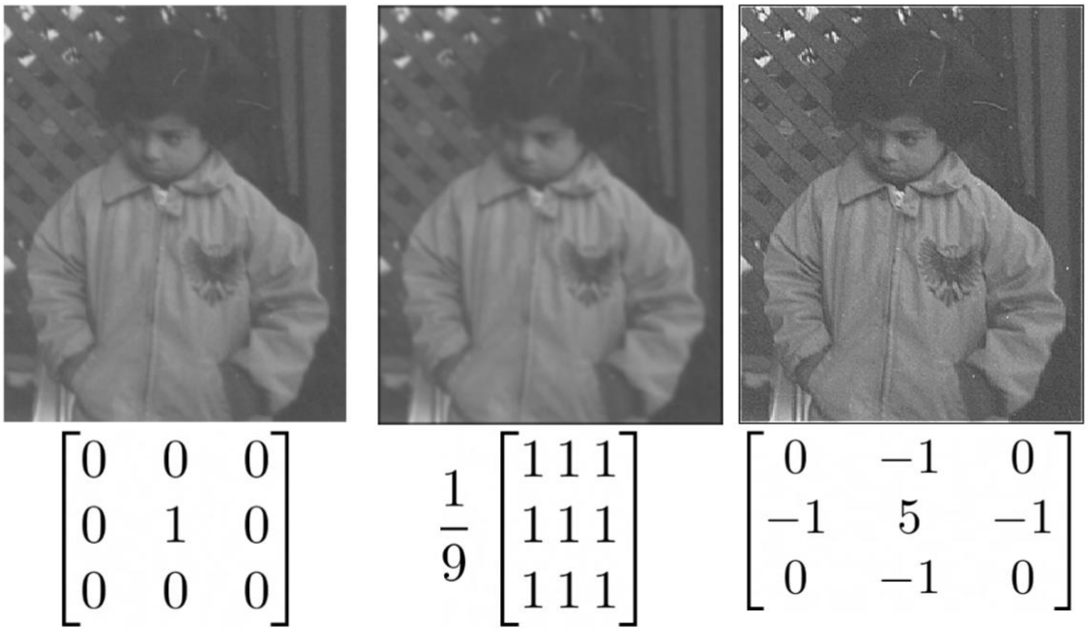

$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$
$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$

Figure 2.9: The top row depicts images convolved with the kernels shown under each image. The leftmost image is the original; when convolved with the identity kernel, the image is unmodified. The middle image was convolved with a 3×3 box blur kernel, and the right image was convolved with a 3×3 sharpening filter.

2.3.3 Convolutional Neural Networks

CNNs are a type of NN intended for application to image data. As the name suggests, they are based on the convolution operation which was discussed in the previous section. In a CNN, the parameters learned throughout network training are convolutional filter values. CNNs can be used to solve multiple types of problems including the following:

- Classification: assign one or more labels to an entire image.
- Object detection or localization: locate objects of interest within an image.
- Semantic segmentation: pixel-wise classification of an image.

CNN Building Blocks

The architecture of a CNN depends on the application it aims to achieve. The focus of this section will be on the building blocks found in CNN architectures for semantic segmentation. Fully connected layers contain nodes where each node is connected to all nodes in the previous layer. Convolutional layers are where convolutional operations take place and are often followed by ReLU ($f(x) = 0\ for\ x < 0, x\ for\ x \geq 0$) or a similar activation function. Once a convolutional kernel is applied, the intermediate result is referred to as a feature map. To interpret what is being learned at layers within a NN, feature maps may be visualized. In some cases, a stride greater than one is used in convolutional layers to achieve downsampling. Max pooling is another technique which achieves downsampling by selecting the maximum pixel value within an area corresponding to a kernel size. For example, if 2×2 max pooling is applied to a 4×4 image with stride 2, the result is a 2×2 image where each pixel corresponds to the largest pixel value in each 2×2 section of the original image (Figure 2.10).

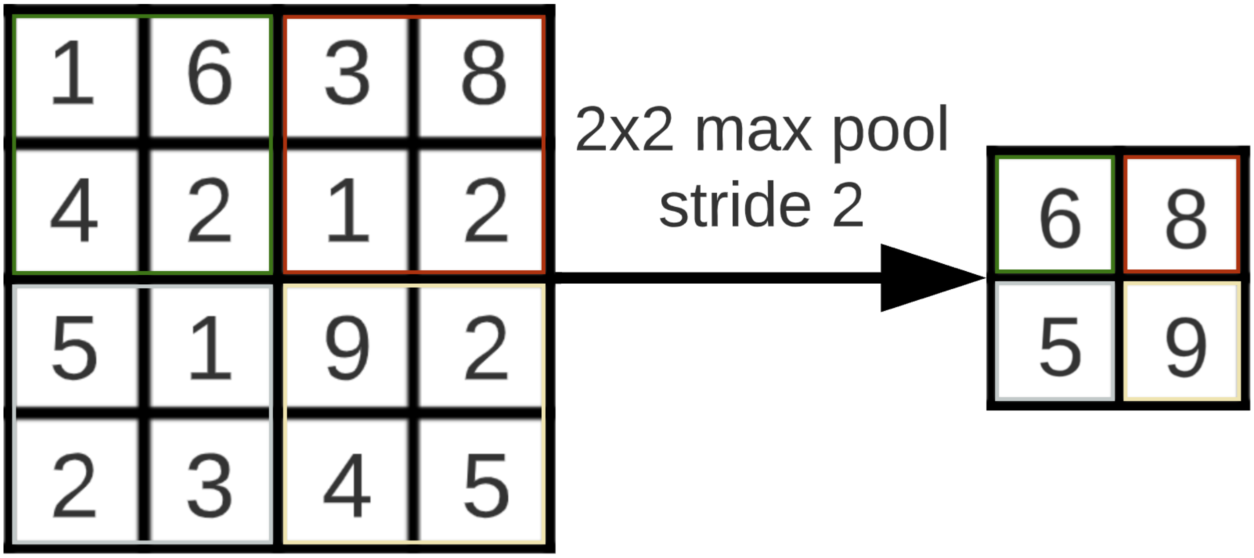


Figure 2.10: Demonstration of max pooling.

A convolutional layer with a 1×1 filter can be used in CNNs to reduce the feature or channel depth [20]. These 1×1 convolutional filters match the input image in channel depth, and the number of 1×1 filters applied corresponds to the feature depth of the convolved feature map. Resampling techniques as previously described can also be implemented in a CNN, such as bilinear upsampling of feature maps to increase their spatial resolution. Transposed convolutions use learnable convolution parameters to upsample feature maps. Skip connections are paths in neural networks which combine feature maps from early layers with those from layers deeper in the network. Skip connections allow information present in early layers of the network to be passed to later layers; this information may have been lost during convolution and resampling operations. Skip connections also provide a shorter path through the network for gradients to travel along during backpropagation. Skip connections alleviate the vanishing gradient problem, in which updates to network parameters become tiny for early layers. Skip connections either involve concatenation or addition; in either case, the feature maps combined with a skip connection should share the same dimensionality. In the case of concatenative skip connections, the feature channels are combined. For example, if two 64×256×256 feature maps are combined with a concatenative skip connection, the result is a 128×256×256 feature map. In additive skip connections, element-wise addition takes place, which results in a feature map with the same shape as the inputs; if two 64×256×256 feature maps are combined with an additive skip connection, the result is also a 64×256×256 feature map.

Patch-Based CNN

Patch-based techniques have seen extensive use in biomedical CNN applications because 3D medical images are often too large to feed into a NN in their entirety [21]. The maximum size of an image which can be fed into a NN is determined by network architecture and hardware memory. Patch-based training involves feeding only a subset (patch) of each scan into the NN at each iteration. For example, if a dataset contains 3D scans with a size of 600×600×600, a network may only consider a 256×256×256 patch of the original scan. At each training and validation iteration, a patch sampling method is used to select the image patch which will be fed into the network. The patch may be selected randomly, weighted towards specific classes, or evenly sampled. In even sampling, a patch has an equal probability of being centered around a voxel corresponding to any label within the scan. In a full body CT dataset with classes of background, heart, brain, and lungs, even sampling would result in a 25% chance of a patch being centered around a voxel of each class. Even sampling allows minority classes to be adequately represented in network training. Patch sampling is reapplied for each scan in every epoch, allowing the network to learn from multiple locations within each example. When a full scan is fed into a network to generate a prediction, a sliding window approach is used. The patch is moved across the image until the entirety of the scan volume is covered. Patch edge issues frequently occur in which segmentations may appear discontinuous at locations where adjacent patches meet. Patch overlap is used to mitigate patch edge issues by allowing adjacent patches to share voxels, and averaging the values of those shared voxels between patches. Patch overlap comes at a computational cost, as increased patch overlap means more patches must be fed through the network [22].

2.3.4 Transfer Learning

Transfer learning is a ML technique which involves the refinement of a model which was trained on a dataset to achieve a related but modified task. When training a NN from scratch, the weights are initialized to small values close to zero, and may follow a distribution dependent on the activation function or number of neurons connected to a given node. Rather than using a random distribution of values as initial weights, transfer learning involves the use of weights which were learned during another training. Transfer learning is thought to be effective because similar tasks are involved in many problems; for example, edge detection and noise reduction may be useful in any computer vision problem. The new network is able to avoid learning these fundamental operations which speeds up model training. Any subset of original model weights can be copied to the new network, meaning it is not necessary that the models share identical architectures.

2.4 Automated Segmentation of the Inner Ear and Round Window

Various CNN and atlas-based approaches have been used to automate the segmentation of various IE structures. Although numerous studies have focused on segmentation of the IE, clear gaps in the literature exist which motivate the current study. This section discusses work related to automated segmentation of IE structures.

Heutink et al. used 2D CNNs to detect and segment the cochlea in ultra-high-resolution CT (UHR-CT) clinical scans and achieved a Dice score of 0.9 [23]. Scans spanned a 4cm cross section of the temporal bone including the IE, suggesting manual preprocessing of scans would be necessary to apply the model to scans with a larger field of view. The approach segmented only the cochlea portion of the IE, not the vestibule or SCCs. A limitation of the study is that all study data originates from two UHR-CT scanners. UHR-CT is also not readily available at many clinical institutions.

Hussain et al. [24] used three cascaded 2D CNNs to segment the full IE in micro-CT scans. A Dice score of 0.9 was achieved by their best model which used the U-net architecture. Each of the three CNNs corresponded to an anatomical plane (axial, coronal, sagittal) to incorporate 3D information to network predictions. All data used in the study originated from two scanners. The study also used cadaveric specimens, as is necessary for micro-CT scan acquisition. Micro-CT scans produce higher quality images than techniques which are applied in clinical practice.

Powell et al. [25] used an atlas-based approach to segment cochlear microstructures in CBCT. Segmented structures include the scala tympani, scala vestibuli, modiolus, and RW. The vestibule and SCCs were not segmented. Dice scores of 0.77 and 0.74 were achieved for the scala tympani and scala vestibuli respectively, and the mean RW centroid distance error was 0.32mm. Ground truth segmentations were based on micro-slicing which involved fixed and dehydrated samples. The approach was additionally validated on 11 clinical CBCT scans from one scanner.

Zhang et al. [26] applied a 3D CNN with a U-net architecture to segment the scala tympani, scala vestibuli, and modiolus. The vestibule and SCCs were not segmented. The model was validated with samples which had both micro-CT and clinical acquisitions. Dice scores of 0.87 and 0.86 were achieved for the scala tympani and scala vestibuli respectively. The scans were manually cropped to a region of interest extending 5mm beyond the boundary of the cochlea, necessitating some manual preprocessing for model use.

The present study is the first which has achieved automated segmentation of the IE and RW in electrode-implanted samples. Dice score and centroid distance results indicate competitive or improved results compared to other work. Additionally, the models presented in this thesis have been validated with extensive multi-institutional clinical data as well as high resolution acquisitions including micro-CT and SR-PCI. Finally, this study was validated by comparing the automated segmentation approach to manual expert segmentations completed by four experts.

References

[1] A. Virzì *et al.*, “Comprehensive Review of 3D Segmentation Software Tools for MRI Usable for Pelvic Surgery Planning,” *J. Digit. Imaging*, vol. 33, no. 1, pp. 99–110, Feb.

2020, doi: 10.1007/s10278-019-00239-7.

[2] L. G. Brown, “A survey of image registration techniques,” *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, Dec. 1992, doi: 10.1145/146370.146374.

[3] R. Jackler and C. Gralapp, “Surgical Anatomy of the Ear,” *Oto Surgery Atlas*. https://otosurgeryatlas.stanford.edu/otologic-surgery-atlas/surgical-anatomy-of-the-ear/ (accessed Feb. 23, 2022).

[4] X. Zhang and R. Z. Gan, “Dynamic Properties of Human Round Window Membrane in Auditory Frequencies,” *Med. Eng. Phys.*, vol. 35, no. 3, pp. 310–318, Mar. 2013, doi: 10.1016/j.medengphy.2012.05.003.

[5] “World report on hearing.” https://www.who.int/publications/i/item/world-report-on-hearing (accessed Feb. 23, 2022).

[6] L. L. Cunningham and D. L. Tucci, “Hearing Loss in Adults,” *N. Engl. J. Med.*, vol. 377, no. 25, pp. 2465–2473, Dec. 2017, doi: 10.1056/NEJMra1616601.

[7] J. P. Roche and M. R. Hansen, “On the Horizon: Cochlear implant technology,” *Otolaryngol. Clin. North Am.*, vol. 48, no. 6, pp. 1097–1116, Dec. 2015, doi: 10.1016/j.otc.2015.07.009.

[8] M. Puechmaille *et al.*, “The French National Cochlear Implant Registry (EPIIC): Bilateral cochlear implantation,” *Eur. Ann. Otorhinolaryngol. Head Neck Dis.*, vol. 137 Suppl 1, pp. S51–S56, Sep. 2020, doi: 10.1016/j.anorl.2020.07.005.

[9] M. Risoud *et al.*, “Sound source localization,” *Eur. Ann. Otorhinolaryngol. Head Neck Dis.*, vol. 135, no. 4, pp. 259–264, Aug. 2018, doi: 10.1016/j.anorl.2018.04.009.

[10] M. S. Harris *et al.*, “Real-Time Intracochlear Electrocochleography Obtained Directly Through a Cochlear Implant,” *Otol. Neurotol. Off. Publ. Am. Otol. Soc. Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol.*, vol. 38, no. 6, pp. e107–e113, Jul. 2017, doi: 10.1097/MAO.0000000000001425.

[11] J.-S. Kim, “Electrocochleography in Cochlear Implant Users with Residual Acoustic Hearing: A Systematic Review,” *Int. J. Environ. Res. Public. Health*, vol. 17, no. 19, p. E7043, Sep. 2020, doi: 10.3390/ijerph17197043.

[12] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, “Medical image registration,” p. 46.

[13] K. A. Miles, “PET-CT in oncology: making the most of CT,” *Cancer Imaging Off. Publ. Int. Cancer Imaging Soc.*, vol. 8 Spec No A, pp. S87-93, Oct. 2008, doi: 10.1102/1470-7330.2008.9015.

[14] S. Sharma, S. Sharma, U. Scholar, and A. Athaiya, “ACTIVATION FUNCTIONS IN NEURAL NETWORKS,” vol. 4, no. 12, p. 7, 2020.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” *ArXiv170802002 Cs*, Feb. 2018, Accessed: Nov. 30, 2021. [Online]. Available: http://arxiv.org/abs/1708.02002

[16] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, “How Does Batch Normalization Help Optimization?,” *ArXiv180511604 Cs Stat*, Apr. 2019, Accessed: Feb. 17, 2022. [Online]. Available: http://arxiv.org/abs/1805.11604

[17] Y. Wu and K. He, “Group Normalization,” *ArXiv180308494 Cs*, Jun. 2018, Accessed: Dec. 07, 2021. [Online]. Available: http://arxiv.org/abs/1803.08494

[18] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *ArXiv150203167 Cs*, Mar. 2015, Accessed: Feb. 17, 2022. [Online]. Available: http://arxiv.org/abs/1502.03167

[19] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J. Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[20] M. Lin, Q. Chen, and S. Yan, “Network In Network,” *ArXiv13124400 Cs*, Mar. 2014, Accessed: Feb. 17, 2022. [Online]. Available: http://arxiv.org/abs/1312.4400

[21] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, “Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification,” *ArXiv150407947 Cs*, Mar. 2016, Accessed: Feb. 24, 2022. [Online]. Available: http://arxiv.org/abs/1504.07947

[22] S. Nikan *et al.*, “PWD-3DNet: A Deep Learning-Based Fully-Automated Segmentation of Multiple Structures on Temporal Bone CT Scans,” *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.*, vol. 30, pp. 739–753, 2021, doi: 10.1109/TIP.2020.3038363.

[23] F. Heutink *et al.*, “Multi-Scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution CT images,” *Comput. Methods Programs Biomed.*, vol. 191, p. 105387, Jul. 2020, doi: 10.1016/j.cmpb.2020.105387.

[24] R. Hussain, A. Lalande, K. B. Girum, C. Guigou, and A. Bozorg Grayeli, “Automatic segmentation of inner ear on CT-scan using auto-context convolutional neural network,” *Sci. Rep.*, vol. 11, no. 1, p. 4406, Feb. 2021, doi: 10.1038/s41598-021-83955-x.

[25] K. A. Powell *et al.*, “Atlas-based segmentation of cochlear microstructures in cone beam CT,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 3, pp. 363–373, Mar. 2021, doi:

10.1007/s11548-020-02304-x.

[26] D. Zhang, R. Banalagay, J. Wang, Y. Zhao, J. H. Noble, and B. M. Dawant, "Two-level Training of a 3d U-Net for Accurate Segmentation of the Intra-cochlear Anatomy in Head CTs with Limited Ground Truth Training Data," *Proc. SPIE-- Int. Soc. Opt. Eng.*, vol. 10949, p. 1094907, Feb. 2019, doi: 10.1117/12.2512529.

Chapter 3

A Deep Learning Solution for Automated Segmentation of the Inner Ear and Round Window on Electrode-Implanted and Non-Implanted Computed Tomography Scans

3.1 Introduction

The cochlea and RW are highly variable IE structures which serve essential roles in hearing [1]–[3]. A highly detailed segmentation of the IE is necessary to trace the lateral wall of the cochlea, which allows cochlear linear and angular length to be determined. An accurate estimate of lateral wall length based on clinical CT is a prerequisite for patient-specific electrode selection [4]. Implantation of an electrode which matches patient anatomy allows the apical region of the cochlea to be stimulated. Stimulation of the apical region in the cochlea, which interprets low frequency sounds, improves hearing outcomes on objective hearing tests [5]–[7]. The RW is an especially important structure of the IE as it marks the entry point for most CI procedures, and is the landmark used as a starting point to measure cochlear duct length and angular depth within the cochlea [8]; these measurements are important for preoperative electrode selection and postoperative customized pitch-mapping.

The goal of the present study was to quickly and accurately segment the IE and RW in clinical CT scans with and without implanted electrodes. This study presents an automated ML pipeline involving two patch-based 3D U-nets with residual connections. The novel pipeline was found to result in accurate segmentations.

3.2 Materials and methods

3.2.1 Image Acquisition

Table 3.1 summarizes all data used throughout the study. Ethics for cadaveric scans were obtained (Ethics # 09052019) (Appendix A). Ethics for clinical scans were obtained through the temporal bone database at the Auditory Biophysics Laboratory (Ethics #112296). Both cadaveric and non-cadaveric CTs were used in the development of the automated segmentation pipeline. Cadaveric scans differ from non-cadaveric scans due to cadaveric sample preparation steps such as extraction of the region of interest, and changes which occur in the body post-mortem. Evaluating the performance of the pipeline on non-cadaveric data ensured the pipeline was suitable for implementation to clinical practice. The pipeline was tested on cadaveric samples because both clinical quality scans and high-quality imaging modalities such as micro-CT and SR-PCI could be acquired for those samples. Micro-CT allows for clear delineation between cochlear turns which is often not visible in clinical modalities such as CBCT or helical CT, and SR-PCI provides exceptional soft tissue contrast [9]. Micro-CT and SR-PCI scans were manually segmented and aligned to clinical modality scans of the same samples. The formation of ground truth labels by segmenting high quality scans allowed for an improved ability to determine pipeline accuracy. Micro-CT and SR-PCI techniques can not be applied to live subjects due to high radiation exposure and limited detector sizes which necessitate dissection of the temporal bone surrounding the IE. Figure 3.1 depicts a qualitative side-by-side comparison of a variety of imaging modalities used in the study. Clinical scans without high resolution ground truth segmentations were segmented directly and used to train the models to increase the dataset size and promote robustness, allowing the models to perform well on scans from a variety of institutions which have differing resolutions and acquisition protocols. CTs were also acquired with and without implanted electrodes to create pipelines which are applicable to scans regardless of the presence of a CI.

Table 3.1: Description of study data. ‘Sample Details’ lists data subsets which share image acquisition modalities. Each bullet point in ‘Input Acquisitions’ represents one acquisition and specifies if an electrode was implanted in the corresponding scans. Acquisitions used in the test set appear in bold. ‘Ground Truth Modality’ and ‘Ground Truth Segmentation Method’ describe the imaging modality and method used to form ground truth segmentations in the respective data subset. The number of samples assigned to train, validation, and test sets are specified in ‘Data Split’.

Sample Details	Input Acquisitions	Ground Truth Modality	Ground Truth Segmentation Method	Data Split
67 cadaveric	● **CBCT 300um isotropic** ● Helical CT 488um×488um×625um Soft Tissue Protocol ● Helical CT 488um×488um×625um Temporal Bone protocol ● Helical CT 488um×488um×1250um Soft Tissue Protocol ● Helical CT 488um×488um×1250um Temporal Bone protocol	Micro-CT 50um isotropic	IE: Micro-CT 50um isotropic manually segmented RW: SR-PCI atlas	Train: 44 Validation: 20 Test: 3
3 cadaveric	**Helical CT 234×234×625um**	● SR-PCI 9um isotropic ● Micro-CT 20um isotropic (1 of 3) ● Micro-CT 40um isotropic (2 of 3)	IE: Micro-CT 20um/40um isotropic manually segmented RW: SR-PCI 9um isotropic manually segmented	Test: 3
60 non-cadaveric	CBCT variety of clinical resolutions	CBCT variety of clinical resolutions	IE&RW: Preliminary network inference and manual correction	Train: 36 Validation: 16 Test: 8

			IE (test set samples): Manual segmentation of scan upsampled to 50um isotropic. RW (test set samples): Expert manual centroid selection	
15 non-cadaveric	● CBCT variety of clinical resolutions ● CBCT variety of clinical resolutions, electrode-implanted low dose (7 of 15) ● CBCT variety of clinical resolutions, electrode-implanted higher dose (7 of 15)	Qualitatively best pre-operative CBCT available	IE&RW: Preliminary network inference and manual correction	Train: 9 Validation: 6
14 cadaveric	Helical CT 234×234×625um	Micro-CT 154um isotropic	IE&RW: Preliminary network inference and manual correction	Train: 7 Validation: 7
6 non-cadaveric	● CBCT variety of clinical resolutions ● CBCT variety of clinical resolutions, electrode-implanted	Qualitatively best pre-operative CBCT available	IE&RW: Preliminary network inference and manual correction	Train: 4 Validation: 2
12 cadaveric	● **CBCT 300um isotropic, electrode-implanted** ● Helical CT 625×234×234um,	Micro-CT 154um isotropic	IE&RW: Preliminary network inference	Train: 8 Validation: 2 Test: 2

	electrode-implanted ● Helical CT 625×488×488um Soft Tissue Protocol, electrode-implanted ● Helical CT 625×488×488um Temporal Bone protocol, electrode-implanted ● Helical CT 1.25×488×488um Soft Tissue Protocol electrode-implanted ● Helical CT 1.25×488×488um Temporal Bone protocol, electrode-implanted		and manual correction IE (test set samples): Micro-CT 154um isotropic upsampled to 50um isotropic and manually segmented RW (test set samples): SR-PCI atlas	

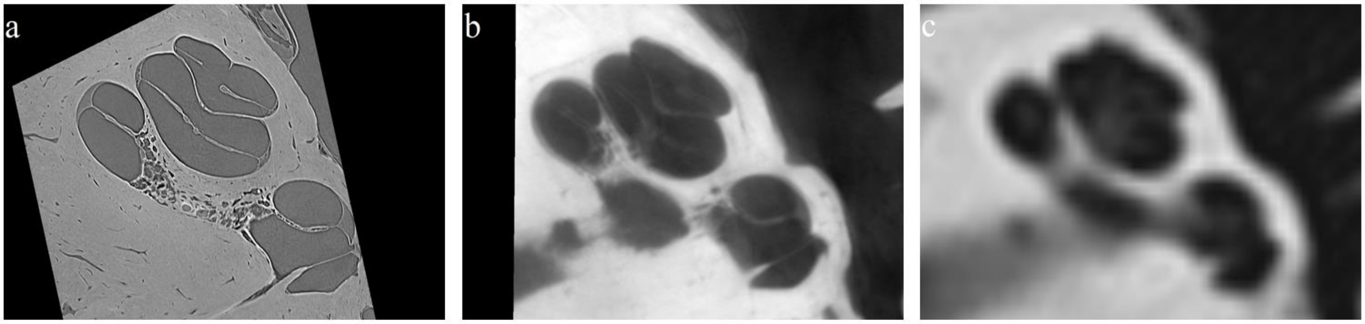


Figure 3.1: Axial slices of varying study data **a:** SR-PCI 9um isotropic, **b:** Micro-CT 40um isotropic, **c:** Helical CT 234×234×625um.

Scanner acquisition settings for non-cadaveric data were not available. Scanner acquisition settings for all cadaveric data are provided. Specimens were scanned by CBCT at varying resolutions using the Xoran xCAT clinical scanner. X-ray tube voltage was 120kV and tube current was 6mA. Specimens were scanned by Helical CT at varying resolutions using the Discovery CT750 HD clinical scanner. X-ray tube voltage was 120kV and tube current was 160mA. Specimens underwent SR-PCI with 9um isotropic voxel spacing at Canadian Light Source Inc. using the Biomedical Imaging and Therapy beamline (05ID-2). Techniques used have been previously published [10], [11]. Specimens were scanned by micro-CT at 20 and 40um isotropic voxel spacing using the GE Locus micro-CT scanner. To fit within the scanner bore, a cylindrical drill bit with a diameter of 40mm and a height of 60mm was used to extract the region of interest from the temporal bone. X-ray tube voltage was 80kV, tube current was 0.45mA, and exposure time was approximately 270 seconds. Specimens were scanned by micro-CT at 50um isotropic voxel spacing using the GE speCZT micro-CT scanner. X-ray tube voltage was 90kV, tube current was 40mA, and exposure time was approximately 4050 seconds. Specimens were scanned by micro-CT at 154um isotropic voxel spacing using the GE Locus Ultra micro-CT scanner. X-ray tube voltage was 120kV and tube current was 20mA.

3.2.2 Data Preparation and Augmentations

Image acquisitions for each sample were rigidly registered. All scans and segmentations included in the study, regardless of original voxel spacing, were resampled to 50um isotropic voxel spacing to promote segmentation smoothness and to ensure a consistent input resolution for CNN training.

Image Segmentation

Two graduate students manually segmented the IE in 67 of the 50um isotropic voxel spacing micro-CT scans. A multi-step manual procedure was used for RW segmentation, as the RW membrane was not easily discernible even in some micro-CT scans. An atlas-based approach was used to segment the RW in the 67 samples with no SR-PCI acquisition. The approach was based on specimens for which SR-PCI data was available. First, rigid landmark registration was performed by selecting points of recognizable anatomy in both the SR-PCI atlas and micro-CT scans such as the tip of modiolus, points on the RW bony overhang, and the stapes footplate. Rigid registration was used in order to maintain the distinctive saddle-like shape of the RW, as affine techniques may result in morphing of the RW to shapes which are not anatomically correct. The SR-PCI atlas was grown by 0.03mm in all directions, smoothed with a 5×5×5 kernel, and resampled to 50um isotropic spacing. A preliminary network with an architecture discussed later (see section 3.2.3) was trained with the 67 samples, the network was preliminary in that it was only trained on a subset of the dataset; this network was applied to further scans to generate preliminary segmentations which were manually corrected. All final segmentations were verified by an expert to ensure consistency and quality control. Figure 3.2 depicts the registration process of clinical and SR-PCI scans as described in the atlas-based approach.

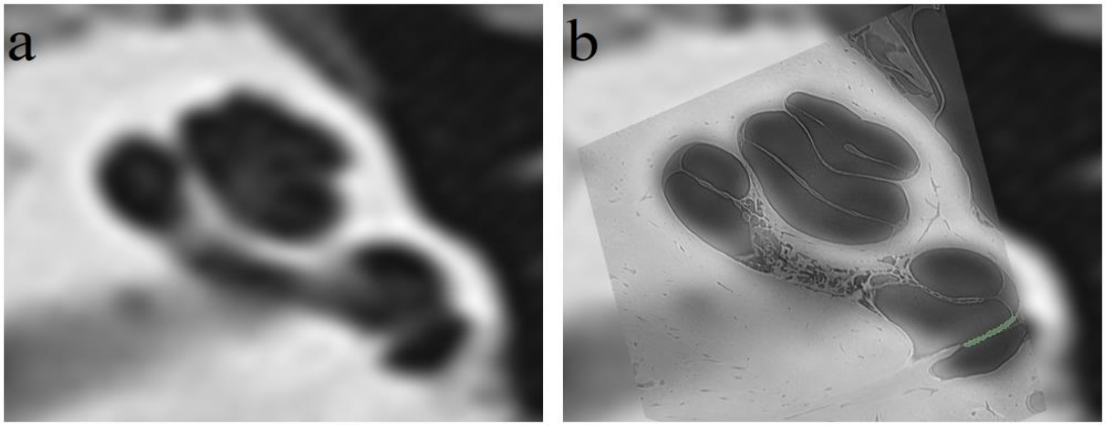


Figure 3.2: 2D visualization of the SR-PCI atlas-based rigid registration. **a:** The fixed image **(**helical CT with 234×234×625um voxel spacing)**, b:** The moving image (SR-PCI 9um isotropic voxel spacing) overlaid upon the fixed image after registration was applied. The resultant RW is displayed in green and identified by an arrow.

As the described atlas-based SR-PCI RW segmentation approach was used to form ground truth segmentations for both training data and some test set scans, validation of the approach was necessary. To evaluate the effectiveness of the method, it was applied to three test set scans which have manual RW segmentations on SR-PCI. The error of the atlas approach was determined by finding the distance between the centroids of the RW produced by the atlas approach and the manually segmented RW based on SR-PCI. The atlas method was applied by registering an SR-PCI atlas to a helical CT with 234×234×625um voxel spacing; as helical CTs have qualitatively inferior imaging characteristics compared to CBCT and micro-CT, the validation demonstrates a worst-case-scenario of the error. RW centroid distances between the manual SR-PCI segmentation and the SR-PCI atlas approach averaged 0.28mm. 3D renderings of the RWs produced by the atlas procedure for the three test set scans are shown in Figure 3.3 alongside the ground truth SR-PCI RW segmentations. As with all atlas-based segmentation approaches, the effectiveness of the method is dependent on how closely the atlas segmentation matches the target segmentation. A case of the atlas poorly matching target anatomy was seen (Figure 3.3b) where the moving atlas appears significantly smaller than the ground truth segmentation for the target scan; this case yielded the largest centroid distance error. Multi-atlas approaches could mitigate atlas mismatch by selecting an atlas with closely corresponding characteristics to the target [12]. Implementation of multi-atlas techniques could lead to improved accuracy of the described RW segmentation protocol.

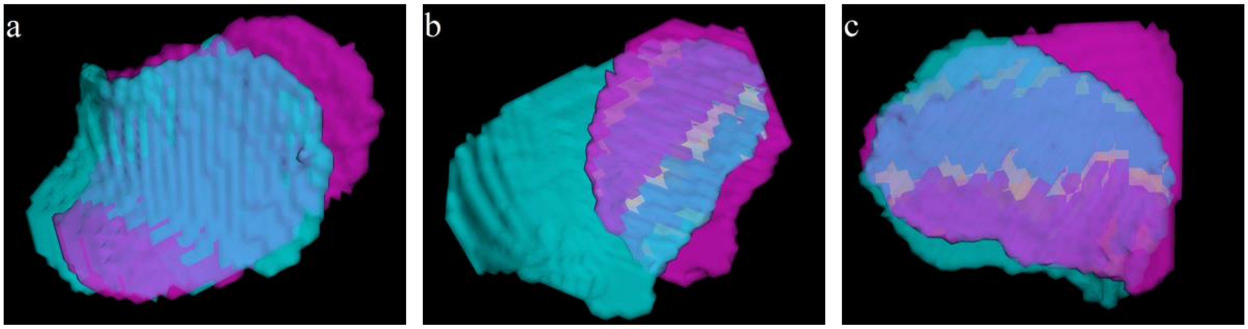


Figure 3.3: 3D visualization of the SR-PCI ground truth RW segmentation in cyan, and the atlas approach RW in fuchsia. **a,c:** atlas fits to target anatomy **b:** atlas differs from target anatomy. 3D visualization was done with 3DSlicer (v4.11, https://www.slicer.org/) [13].

For the three test set samples with both micro-CT (20um or 40um) and SR-PCI acquisitions, the IE was manually segmented based on micro-CT, as the micro-CT field of view included the vestibule and semicircular canals, unlike SR-PCI which had a field of view limited to the cochlea. The RW was segmented based on SR-PCI, grown by 0.03mm in all directions, and smoothed by a 5×5×5 kernel. IE and RW segmentations were combined, and voxels independently assigned to both IE and RW were labeled as RW.

For the eight non-cadaveric test set CBCT scans, the IE was manually segmented based on the CBCTs after they were upsampled to 50um isotropic voxel spacing. Rather than voxel-wise segmentation of the RW, an expert otolaryngologist manually marked the RW centroids in the scans by viewing the samples in axial, sagittal, and coronal view planes.

Four domain experts, including two radiologists and two surgeons, manually segmented eight test set cadaveric CT scans using the Dragonfly (v 2021.1, http://www.theobjects.com/dragonfly) software toolkit with supervision of a graduate student experienced with the software. Six CTs were segmented with no electrode, including three 300um isotropic voxel spacing CBCT scans, three 234×234×625um voxel spacing helical CT scans, and two 300um isotropic voxel spacing CBCT scans with implanted electrodes. These acquisitions were selected to be representative of what could be expected in clinical practice. Scans were manually segmented by experts at clinical resolution to keep segmentation times reasonable. To improve the expert segmentations, each was post-processed by upsampling them to 50um isotropic voxel spacing and Taubin smoothing was applied. Taubin smoothing was introduced to alleviate volume shrinkage, which is a common occurrence among traditional smoothing approaches [14]. Taubin smoothing was selected based on qualitatively preferential results compared to Gaussian and median filtering. Expert segmentations were post-processed to make comparison to the automated pipeline more fair, as the automated pipeline included upsampling and post-processing steps. Figure 3.4 depicts 3D renderings of the expert segmentation before and after post-processing. To confirm that the described post-processing improved expert segmentations, Dice scores were analyzed on a per-expert basis (Figure 3.5). For all experts, the post-processing improved Dice scores. All data used in the following *t* tests were tested for normality with the Kolmogorov-Smirnov test of normality. A paired *t* test was performed to compare Dice scores of expert segmentations to post-processed expert segmentations. The *t* test suggests that there was a statistically significant difference in Dice score between the unprocessed segmentations (mean = 0.846, std dev = 0.031, 95% CI = 0.835 to 0.857) and the processed segmentations (mean = 0.865, std dev = 0.034, 95% CI = 0.853 to 0.877) ($t = 19.28$, $p<0.0001$). All future comparisons of expert segmentations consider only the post-processed versions.

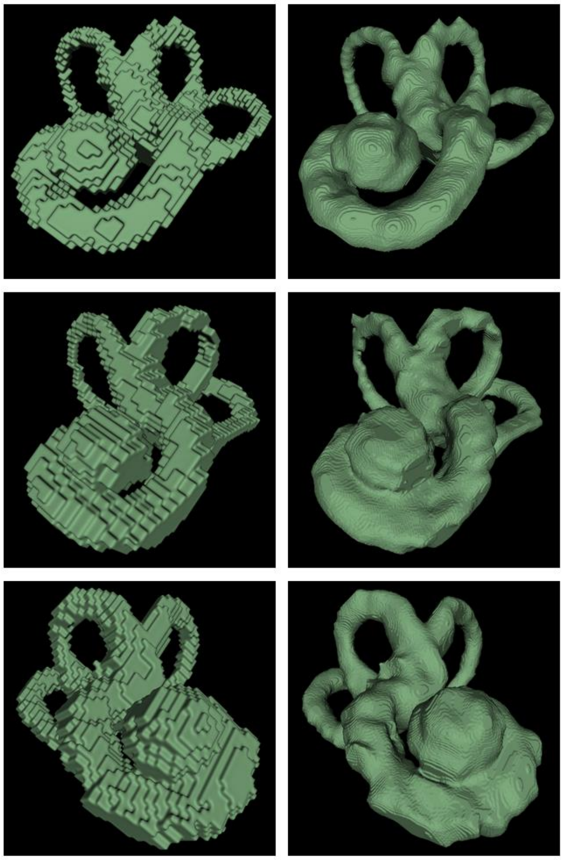


Figure 3.4: Expert segmentation post-processing. The left column depicts unprocessed expert segmentations, and the right column depicts post-processed segmentations. From top to bottom, segmentations correspond to: 300um isotropic voxel spacing CBCT, 234×234×625um helical CT, implanted 300um isotropic voxel spacing CBCT.

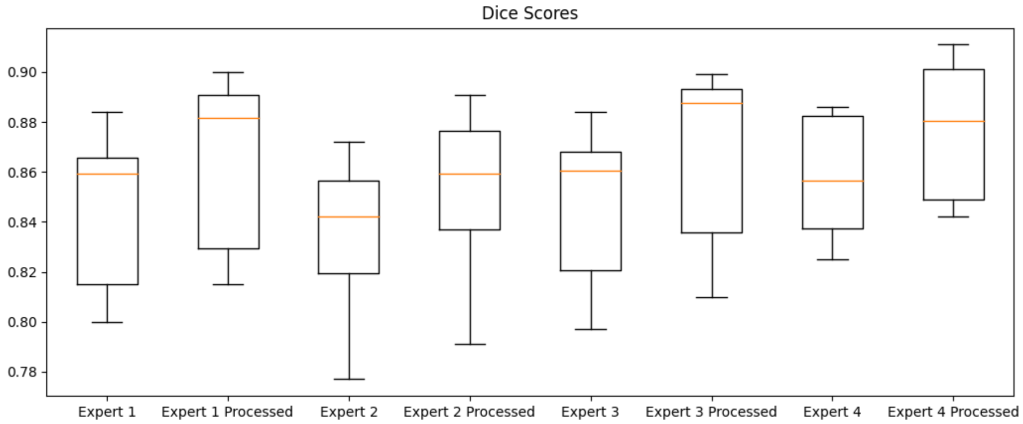


Figure 3.5: Dice scores with and without post-processing.

The simultaneous truth and performance level estimation (STAPLE) algorithm is an iterative approach to combine multiple expert segmentations of one sample [15]. STAPLE estimates the accuracy of experts and weights the contribution of their segmentation to the final consensus segmentation accordingly. Dice scores of the STAPLE consensus are plotted alongside expert scores in Figure 3.6. A paired *t* test was performed to compare the Dice score of processed expert segmentations to STAPLE consensus segmentations. The *t* test suggests that for each expert there was a statistically significant difference in Dice score between the processed segmentations (mean = 0.865, std dev = 0.034, 95% CI = 0.853 to 0.877) and the STAPLE consensus segmentations (mean = 0.882, std dev = 0.031, 95% CI = 0.861 to 0.885) (t = 4.17, p = 0.0002). The STAPLE consensus represents a best-case scenario for expert segmentations, although in clinical practice it would be infeasible to have multiple experts manually segment scans.

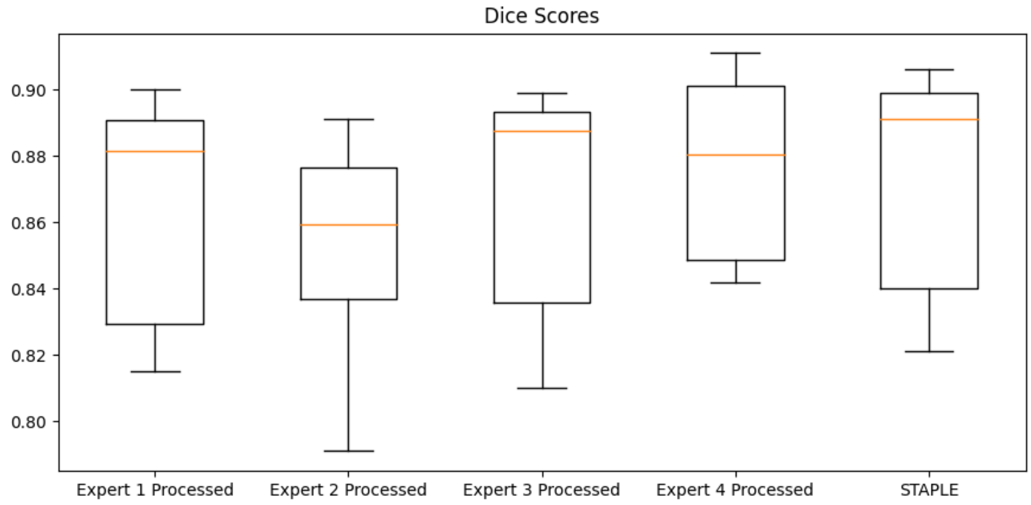


Figure 3.6: Dice scores of processed expert segmentations and STAPLE consensus.

Image Augmentation

Extensive image augmentations were applied to the acquired data. All image augmentations were intended to represent realistic variations of the data, thereby expanding the dataset size and adding variance. All scans were flipped in the sagittal plane, effectively doubling the dataset size as all samples appeared as both a right and left IE. Each scan was modified with two levels of additive Gaussian noise with a mean of zero and standard deviation sampled from a uniform distribution.

Noise standard deviations ranged from 130-240 and 240-350 HU for low and high noise versions, respectively. Figure 3.7 depicts a scan modified with the described levels of additive Gaussian noise. Noise was added to scans at their native resolutions, rather than after the scans were upsampled to 50um isotropic voxel spacing. Adding noise to scans at their native resolution resulted in noise profiles which appeared more realistic, as adding noise after upsampling resulted in noise which was higher resolution than could have been present in the original scans. B-spline interpolation was used to upsample all scans, as other interpolation techniques led to imaging artifacts. Nearest neighbor resampling was used to resample segmentations to 50um isotropic voxel spacing. Scans were resampled to the image grid of their corresponding ground truth segmentations to ensure alignment. All scans were cropped via script to a bounding box 5mm beyond the extent of the IE ground truth segmentation in all directions. Following cropping and upsampling, image volumes averaged 499×483×467 voxels.

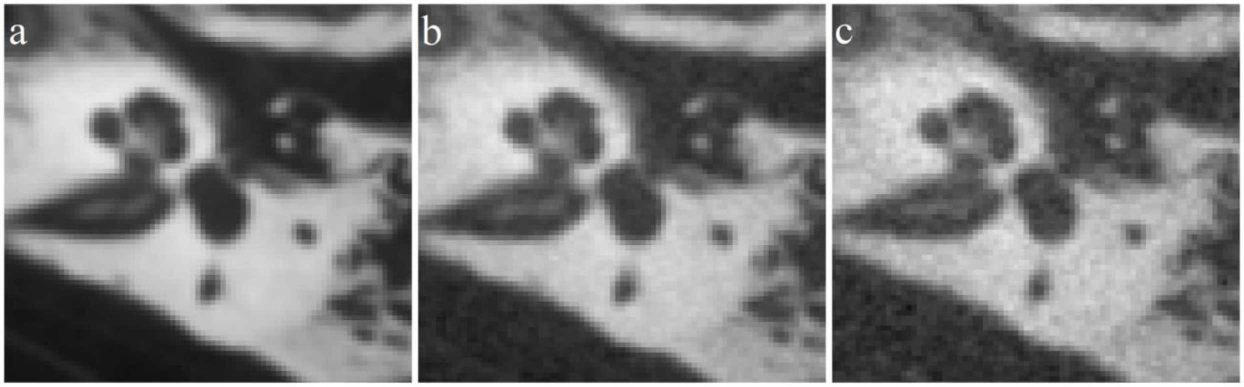


Figure 3.7: Axial view of 300um isotropic cadaveric CBCT scans displayed without interpolation. **a:** scan without artificial noise, **b:** scan with low artificial noise (standard deviation 185 HU), **c:** scan with high artificial noise (standard deviation 295 HU).

To select the standard deviations of noise applied to scans, non-cadaveric scans were analyzed. Non-cadaveric scans often contain more noise compared to cadaveric scans due to the removal of anatomy surrounding the region of interest reducing radiation scatter [16]. By adding noise similar to what was observed in non-cadaveric scans to cadaveric scans, the cadaveric scans became more representative of scans the pipeline could encounter in clinical practice. Qualitatively noisy non-cadaveric scans were identified by eye, and image intensity values were sampled along constant density regions of the scans (Figure 3.8). In a scan acquisition with no noise present, voxel intensities should be relatively constant in regions with consistent density such as portions of the temporal bone or the fluid filled inner auditory canal. Standard deviations of the sampled image intensities were found to be <50 HU for cadaveric samples and >200 HU for clinical scans. A maximum standard deviation of noise applied was selected as 350 HU, to act as an overestimate which accounts for noisier acquisitions not present in the dataset.

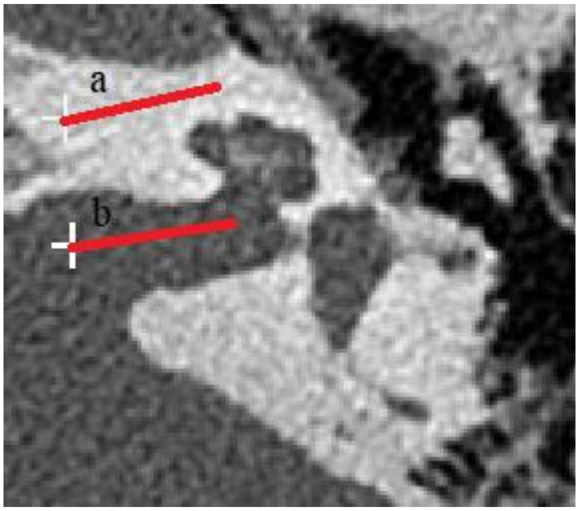


Figure 3.8: Axial view of voxel intensities sampled on a non-cadaveric CBCT scan without interpolation. **a:** a line through the temporal bone near the cochlea, **b:** a line through the fluid filled internal auditory canal.

To determine the effectiveness of the artificial noise augmentation before application to all network training data, preliminary networks with only the first 67 samples were trained with and without the noise augmentations. All other network hyperparameters remained unchanged between the two trainings, meaning all differences in results were attributed to the presence or lack of the artificial noise augmentation. Figure 3.9 demonstrates the superior performance of the preliminary network which was trained with the artificial noise augmentation. Improved continuity in the SCCs, reduced false positives around the inner auditory canal, and improved smoothness in the cochlea were produced by the model trained with the artificial noise augmentation.

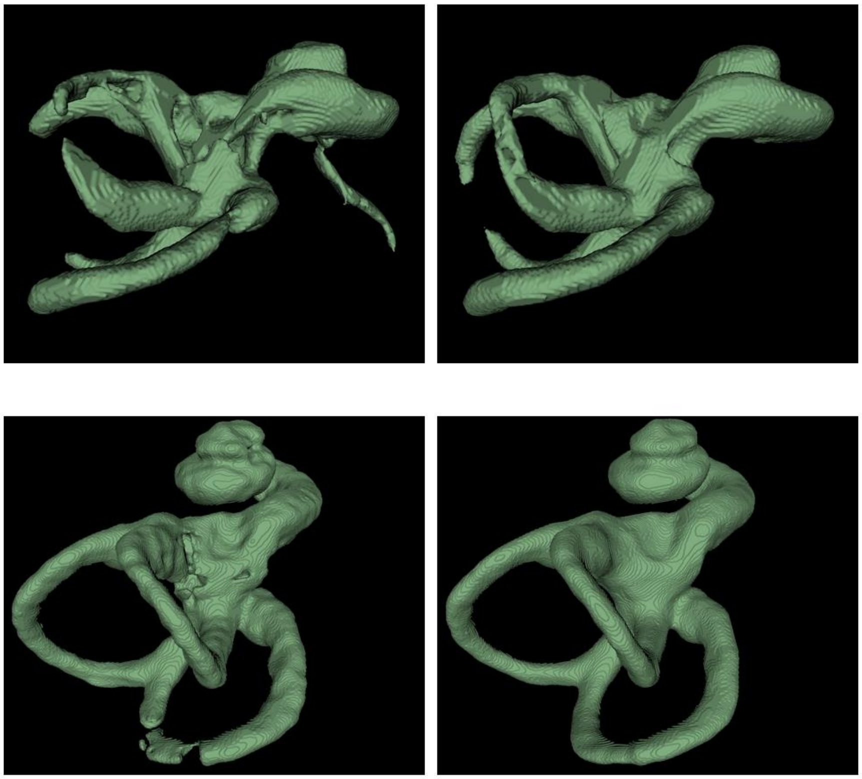


Figure 3.9: 3D renderings of pipeline outputs for two test set non-cadaveric scans. The left column shows inferences from the model trained without noise augmentations. The right column shows inferences from the model with the noise augmentation.

3.2.3 Segmentation Pipeline

Of the 151 samples included in the no-electrode network, 96 were assigned to the training set, 49 to the validation set, and six to the test set. Of the 25 samples included in the electrode-implanted network, 17 were assigned to the training set, six to the validation set, and two to the test set. All image acquisitions of the same sample were assigned to the same set.

The networks were trained using a workstation with a Xeon ES-2640 v4 CPU, 256GB RAM, and an A6000 GPU with 48GB VRAM, which was essential for a large training patch size of 288×256×256. Although theoretical justification has not been formally proven, large patch sizes have seen promising results as they provide increased spatial context to CNNs [17], [18]. The NVIDIA CLARA Train (v 3.1, https://developer.nvidia.com/clara) toolkit was utilized to facilitate model training and inference. The networks were trained with a batch size of one, which allowed patches to span the maximum receptive field size based on VRAM limitations. The ADAM learning rate optimizer was used with a learning rate of 0.0001. The network which achieved the maximum validation Dice score across the IE and RW labels was selected as the best network to prevent overfitting.

The no-electrode network was trained for a total of 235 epochs. The electrode-implanted network was trained for 50 epochs with initial weights set to values learned during the no-electrode network training. Fine-tuning trained networks with data sharing characteristics of the original network has been shown to speed up convergence [19].

The network architecture followed a U-net-like encoder-decoder paradigm (Figure 3.10) with residual skip connections adapted from Myronecko et al. without inclusion of the variational autoencoder portion [20]; a starting point for hyperparameter values were selected in consideration of this paper as well as the IE segmentation papers discussed in section 2.4. Hyperparameters were adjusted one at a time to find ideal settings for the given segmentation task. The model utilized 3D convolutions with 3×3×3 convolutional kernels. On the contracting path, image volumes were progressively downsampled and feature depth was increased. On the expanding path, fewer kernels were used and feature maps were upsampled until the output matched the input size. The output, after being passed through the softmax function, was a three-channel probability map with probabilities indicating the model's certainty that each voxel belonged to classes 0 (background), 1 (IE), or 2 (RW).

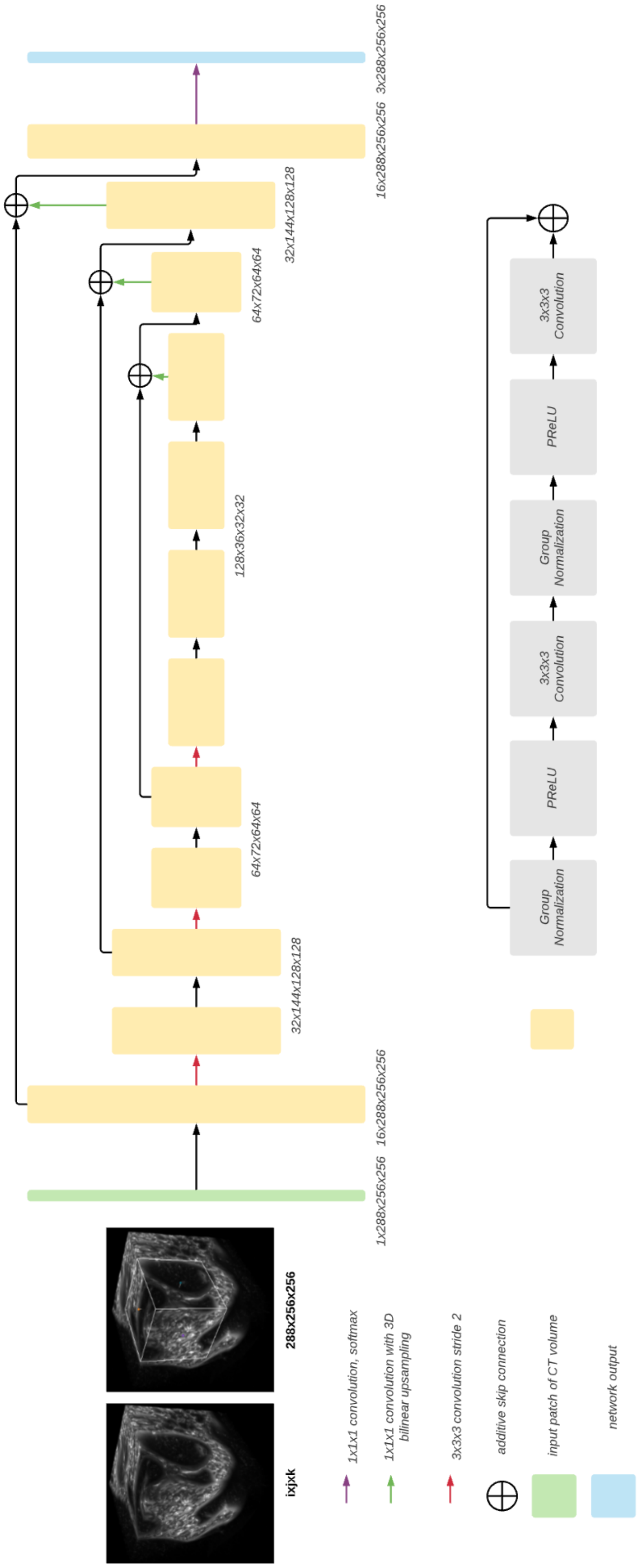


Figure 3.10: CNN architecture for 50um IE and RW segmentation.

Group normalization was utilized in each convolutional block with a group size of eight. Group normalization was selected over batch normalization as it has superior performance on small batch sizes [21]. L2 regularization was applied to all convolutional kernels with a regularization weight of 1e-5. A dropout rate of 0.2 was applied after the first encoder convolution. The PReLU activation function was applied in each convolutional block; PReLU is similar to ReLU, except instead of all values less than zero being set to zero, they are represented by a linear relationship which has a slope that is a learnable parameter. PReLU is thought to be effective as ideal nonlinearity may depend on the depth of the network at which convolutions are being applied.

Voxel intensities were normalized to zero mean and unit standard deviation before being fed into the network. To account for differences in scanner acquisition protocols, image intensities were uniformly shifted during training between -30% and +30% of their original values. This range was selected through analysis of intensity differences in multiple CT modalities on both cadaveric and non-cadaveric data. To determine if a uniform shift of image intensities would be beneficial to network training, mean image intensities for 33 study scans all cropped to a 5mm region surrounding the IE were plotted (Figure 3.11). As these scans were all cropped to a similar region of interest, their intensities were directly comparable. A wide range of average intensities were observed which indicated that shifting image intensities should add realistic variance to the dataset.

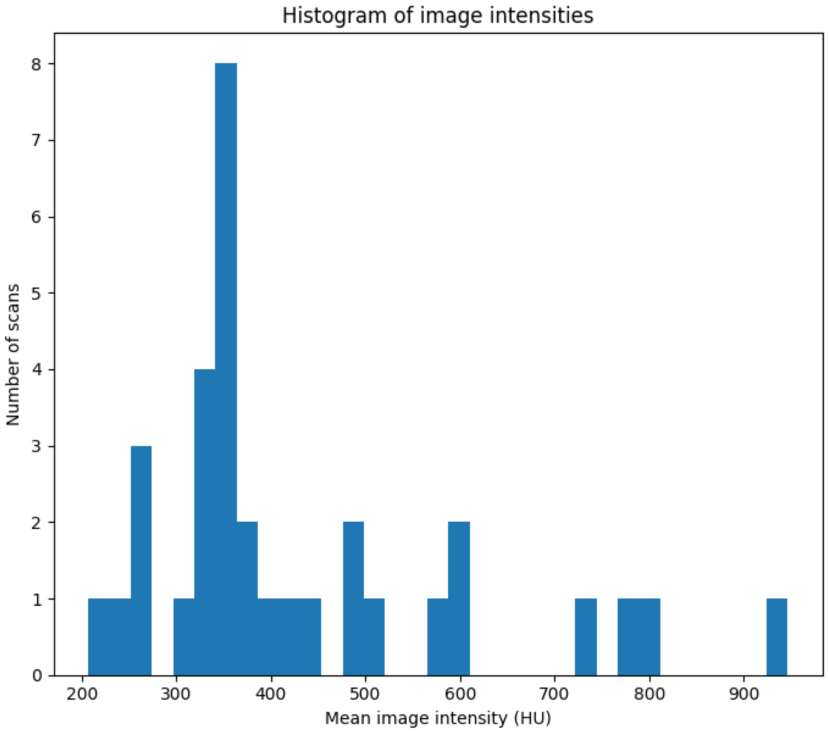


Figure 3.11: Average image intensities for 33 study scans including 17 non-cadaveric CBCT scans, eight cadaveric CBCT scans, and eight cadaveric helical CT scans.

A hybrid loss function representing an equally weighted sum of both focal loss $(-(1-p_t)^{\gamma}log(p_t), \gamma = 2)$ and Dice loss $(1-\frac{2|X \cap Y|}{(|X|+|Y|)})$ was utilized. Dice loss aims to maximize overall segmentation similarity, and focal loss aims to mitigate class imbalance by emphasizing loss contribution of poorly classified voxels [22]. Due to the relative anatomical sizes of the IE and RW, the data showed clear class imbalance as the number of RW voxels was far outnumbered by IE and background voxels (97% background, 3% IE, <0.007% RW). Loss associated with the background class was included in the loss calculation, as excluding background loss contributions led to extensive RW false positives (background voxels incorrectly classified as RW).

A balanced patch sampling technique was used for network iterations in which a single patch from scan volumes was selected with equal probability of being centered around background, IE or RW voxels. Balanced sampling was utilized to ensure a sufficient number of minority-class patches were seen in network training and validation. During inference, a sliding window approach was utilized with 70% patch overlap. Increased overlap minimized patch artifacts by averaging the probabilities of shared voxels in adjacent patches, and resulted in a computational tradeoff as more patches were fed through the network to attain the final segmentation [23].

To complete the pipeline, segmentations produced by the network were post-processed. First, the argmax function was used to convert the three-channel probability map produced by the final softmax activation function into a segmentation label map with values of 0, 1, or 2. Next, Taubin smoothing was applied which removed subtle patch lines introduced by the patch-based inference approach. Island removal was then applied to each structure, which ensured that only the largest island was kept. Pipeline outputs at each stage of post-processing are visualized in Figure 3.12.

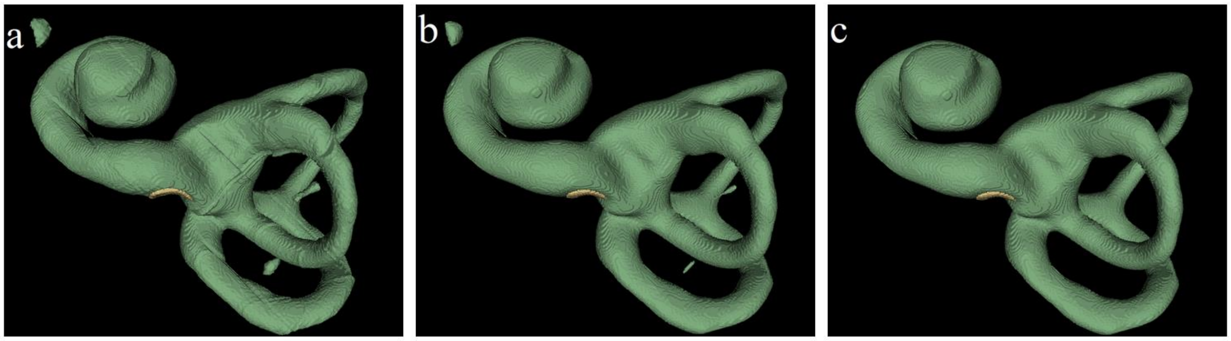


Figure 3.12: 3D renderings of intermediate pipeline segmentations for a non-cadaveric CT. **a:** raw network output. Subtle patch lines are present as a result of the patch-based sliding window inference approach, **b:** the segmentation after Taubin smoothing. Small false positive islands are visible, **c:** the final pipeline output after island removal.

3.2.4 Evaluation Metrics

Numerous metrics were calculated to evaluate the segmentation pipeline. Dice coefficient (DC) ($\frac{2|X \cap Y|}{(|X|+|Y|)}$) described the similarity of ground truth and pipeline output segmentations. Centroid distance (CD) ($||C_1 - C_2||, C_1 = (x_1, y_1, z_1)\,, C_2 = (x_2, y_2, z_2)$) was used to find the distance between the centroid of the ground truth RW and that of the RW predicted by the pipeline. Centroid distance was used for the RW as the Dice coefficient may be a poor measure of accuracy when applied to a thin structure such as the RW. Maximum Hausdorff Distance (MHD)

$(max\{sup_{x\in X}(inf_{y\in Y}d(x,y)), sup_{y\in Y}(inf_{x\in X}d(x,y))\})$ gave an idea of the largest segmentation errors, and was calculated by sampling the pipeline output at all vertices and finding the nearest point, including any vertex, edge, or mesh face on the ground truth segmentation; the largest distance sampled is the MHD. The average Hausdorff Distance (AHD) is the average distance between sampled points when calculating MHD.

3.3 Results

Network inference was tested on an alternate workstation with an i7-7770k CPU, TITAN RTX GPU with 24GB VRAM, and 64GB RAM. This workstation is less powerful than the workstation used to train the models and is more representative of what could be available in a hospital setting. In total, the pipeline took 197 s (upsampling, 1 s; network inference, 166 s; and post-processing, 30 s) to segment the IE and RW with 70% patch overlap.

3.3.1 Testing on Cadaveric Scans with and Without Implanted Electrodes

This section presents a qualitative and quantitative evaluation of the accuracy of the pipeline. The pipeline was tested using two cadaveric CBCT with implanted electrodes as well as three helical CT and three CBCT without implanted electrodes. Pipeline segmentations are evaluated by comparison to ground truth segmentations and manual expert segmentations.

Comparison to Ground Truth Segmentations

Here, ground truth segmentations are compared to pipeline outputs. Dice scores of 0.877 and 0.925 were achieved for the IE in implanted and non-implanted samples, respectively. Distances between predicted and ground truth RW centroids were 0.567mm and 0.178mm in implanted and non-implanted samples, respectively. Table 3.2 depicts further quantitative analysis. Results follow an expected trend based on the characteristics of each test set subset, with no-electrode metrics superior to that of electrode-implanted, and with no-electrode 300um isotropic CBCT metrics superior to no-electrode 234×234×625um helical CT. Figure 3.13 depicts 2D and 3D visualizations of pipeline output segmentations. IE and RW segmentations appear anatomically accurate even in electrode-implanted samples where metallic artifacts directly cover predicted RW segmentations (Figure 3.13g,h).

Table 3.2: Mean values are presented for all metrics. Data ranges are given in parenthesis. Metrics include Inner Ear Dice Coefficient (IE DC), Inner Ear Average Hausdorff Distance (IE AHD), Inner Ear Maximum Hausdorff Distance (IE MHD), and Round Window Centroid Distance (RW CD). Metrics are presented for three test set subsets as each is unique in clinical and ground truth modalities.

Clinical Modality	**IE DC**	**IE AHD (mm)**	**IE MHD (mm)**	**RW CD (mm)**
Helical CT 234×234×625um	0.916 (0.012)	0.069 (0.01)	0.628 (0.284)	0.240 (0.217)
CBCT 300um isotropic	0.923 (0.028)	0.067 (0.031)	0.459 (0.219)	0.116 (0.073)
CBCT 300um isotropic electrode-implanted	0.877 (0.016)	0.102 (0.026)	0.734 (0.613)	0.567 (0.025)

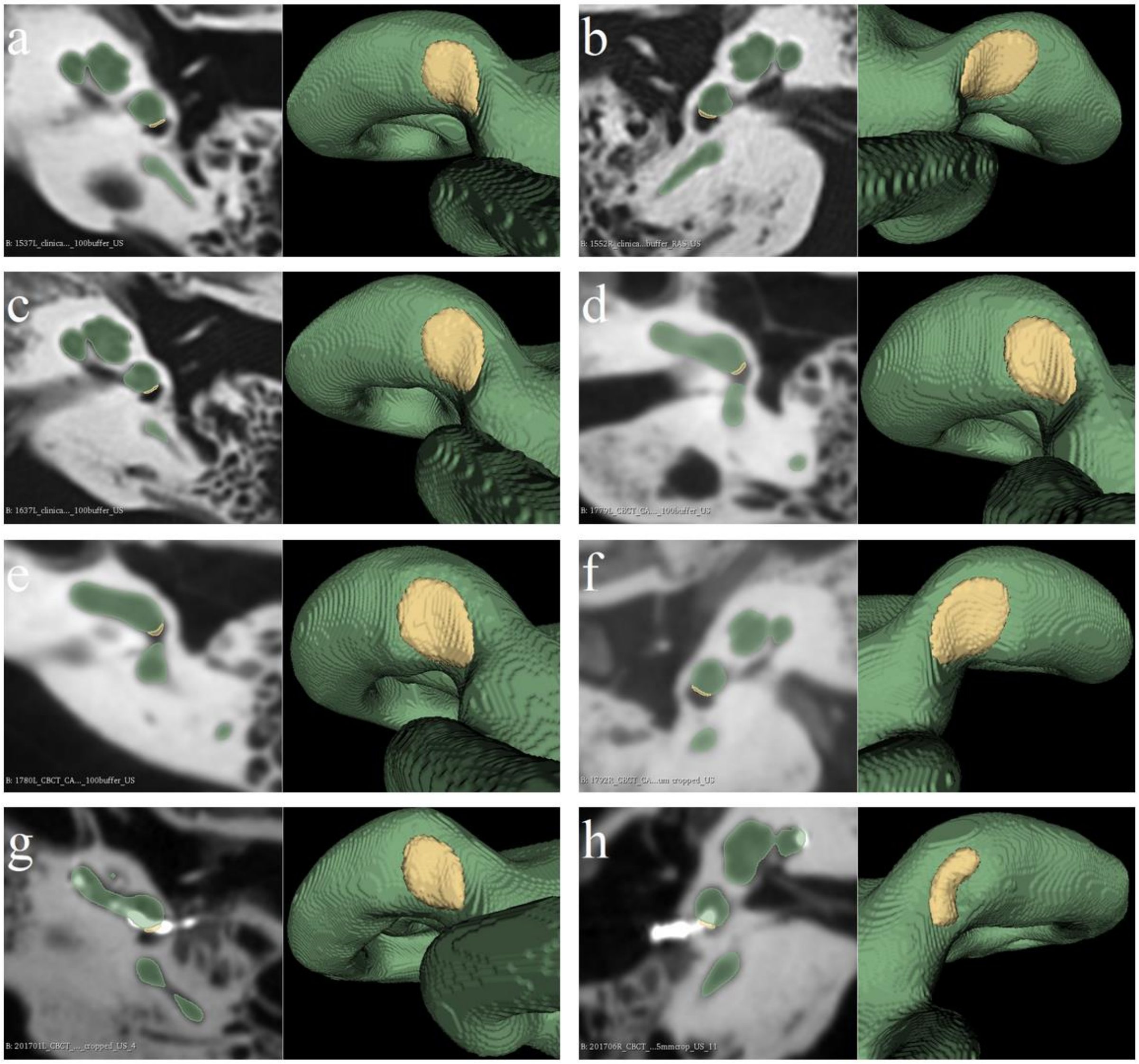

a
B: 1537L_clinica..._100buffer_US
b
B: 1552R_clinica...buffer_RAS_US
c
B: 1637L_clinica..._100buffer_US
d
B: 1779L_CBCT_CA..._100buffer_US
e
B: 1780L_CBCT_CA..._100buffer_US
f
B: 1792R_CBCT_CA....um cropped_US
g
B: 201701L_CBCT_...._cropped_US_4
h
B: 201706R_CBCT_...5mmcrop_US_11

Figure 3.13: 2D and 3D visualizations of pipeline output segmentations. IE (depicted in green) and RW (depicted in yellow) can be seen in all eight test set scans. **a,b,c:** 234×234×625um, **d,e,f:** 300um isotropic, **g,h:** 300um isotropic electrode-implanted.

To interpret the localized accuracy of the segmentation pipeline, a 3D visualization of the shortest Euclidean distances between points on the pipeline output and ground truth were generated (Figure 3.14). The largest segmentation errors, on the order of 0.2-0.35mm, were found in areas where low clinical CT contrast heavily impaired image quality such as the modiolus and between cochlear turns. In the electrode-implanted samples, mismatches of a similar magnitude were present in areas impacted by electrode artifacts. Segmentation errors at the apex of the cochlea, which marks the endpoint for cochlear duct measurements, were on the order of 0.05mm.

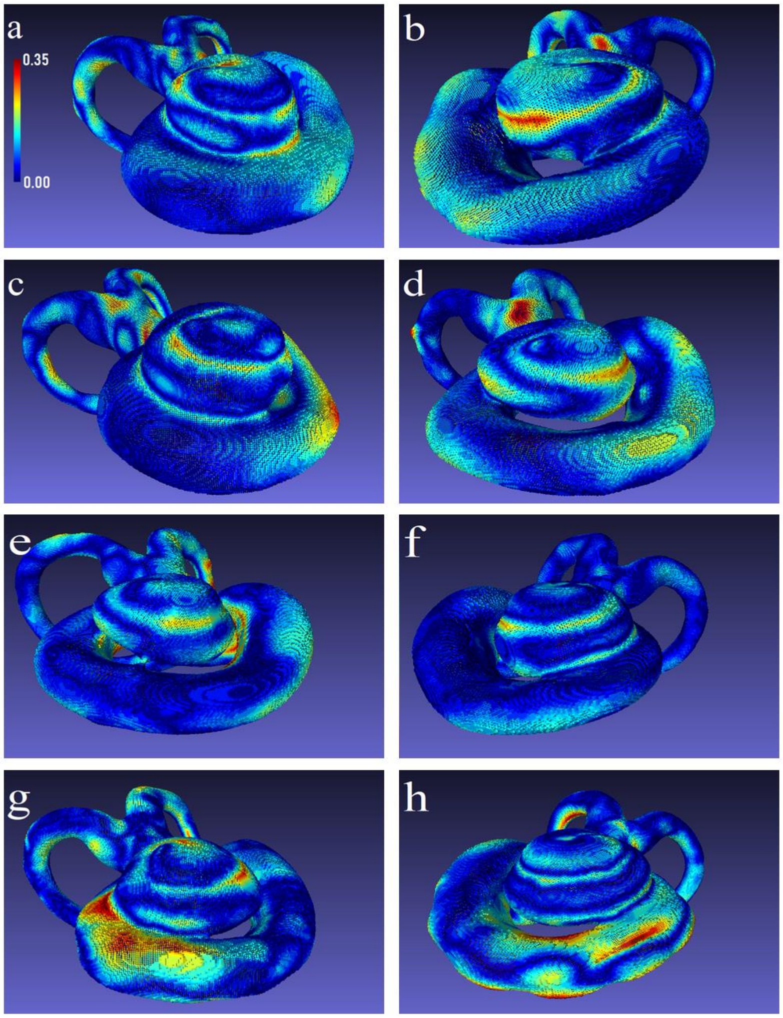


Figure 3.14: Distances between the pipeline output and ground truth segmentations are depicted by coloured pipeline output segmentations for eight test set scans. Distances are in units of mm and follow the scale shown in image **a**. **a,b,c:** 234×234×625um, **d,e,f:** 300um isotropic, **g,h:** 300um isotropic electrode-implanted. 3D renderings and distance calculations were completed with MeshLab [24] (v 1.3.2, https://www.meshlab.net/).

Comparison to Expert Segmentations

Here, processed expert segmentations of clinical resolution scans are compared to automated pipeline segmentations. Figure 3.15 shows that pipeline Dice scores are superior to that of processed manual expert segmentations, and that pipeline segmentation quality is more consistent than manual segmentations. As Figure 3.15 depicts results across all eight test set samples, the outlier related to the pipeline corresponds to an implanted sample for which poorer model performance is expected.

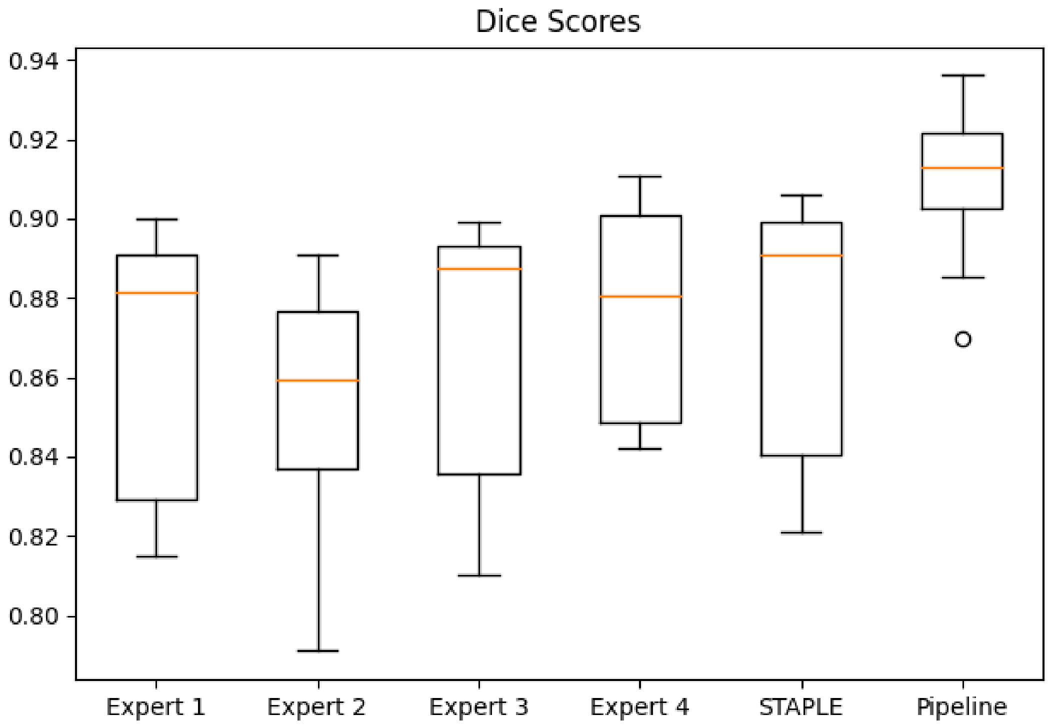


Figure 3.15: Dice scores comparing processed expert segmentations, STAPLE consensus, and the automated pipeline.

All data used in the following ANOVA and *t* tests were tested for normality with the Kolmogorov-Smirnov test of normality. A one-way repeated measures ANOVA test was performed to evaluate how Dice score performance varied by expert. The one-way ANOVA suggests that there was not a statistically significant difference in Dice score between expert segmenters (F = 2.98, p = 0.0837). A paired *t* test was performed to compare Dice score of expert segmenters (grouped together) and the automated pipeline. The *t* test suggests that there was a statistically significant difference in Dice score between the experts (mean = 0.865, std dev = 0.034, 95% CI = 0.853 to 0.877) and the automated pipeline (mean = 0.909, std dev = 0.021, 95% CI = 0.902 to 0.916) (t =

10.04, p<0.0001). A paired *t* test was performed to compare STAPLE consensus segmentations and the automated pipeline. The *t* test suggests that there was a statistically significant difference in Dice score between the STAPLE consensus (mean = 0.873, std dev = 0.035, 95% CI 0.849 to 0.897) and the automated pipeline (mean = 0.909, std dev = 0.022, 95% CI 0.894 to 0.924) (t = 4.02, p = 0.0051). These results indicate that the pipeline outperforms both processed expert segmentations, and an expert consensus segmentation generated by STAPLE. The Bonferroni correction indicates that all Dice comparisons included in the thesis are statistically significant with $\alpha = 0.05$. As the test set was composed of three subsets (300um isotropic CBCT with implanted electrodes, 234×234×625um Helical CT, and 300um isotropic CBCT), processed expert segmentations were compared to the automated pipeline on a per test set subset basis in Figure 3.16. Processed expert segmentations are visualized alongside ground truth segmentations and pipeline outputs in Figure 3.17.

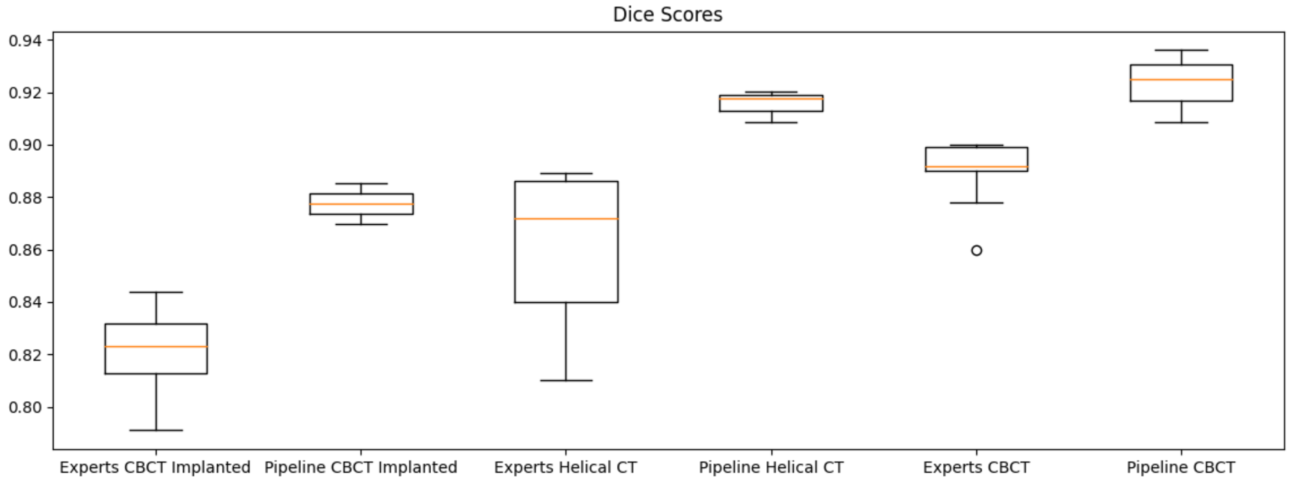


Figure 3.16: Dice scores comparing processed expert segmentations to pipeline segmentations based on each test set data subset.

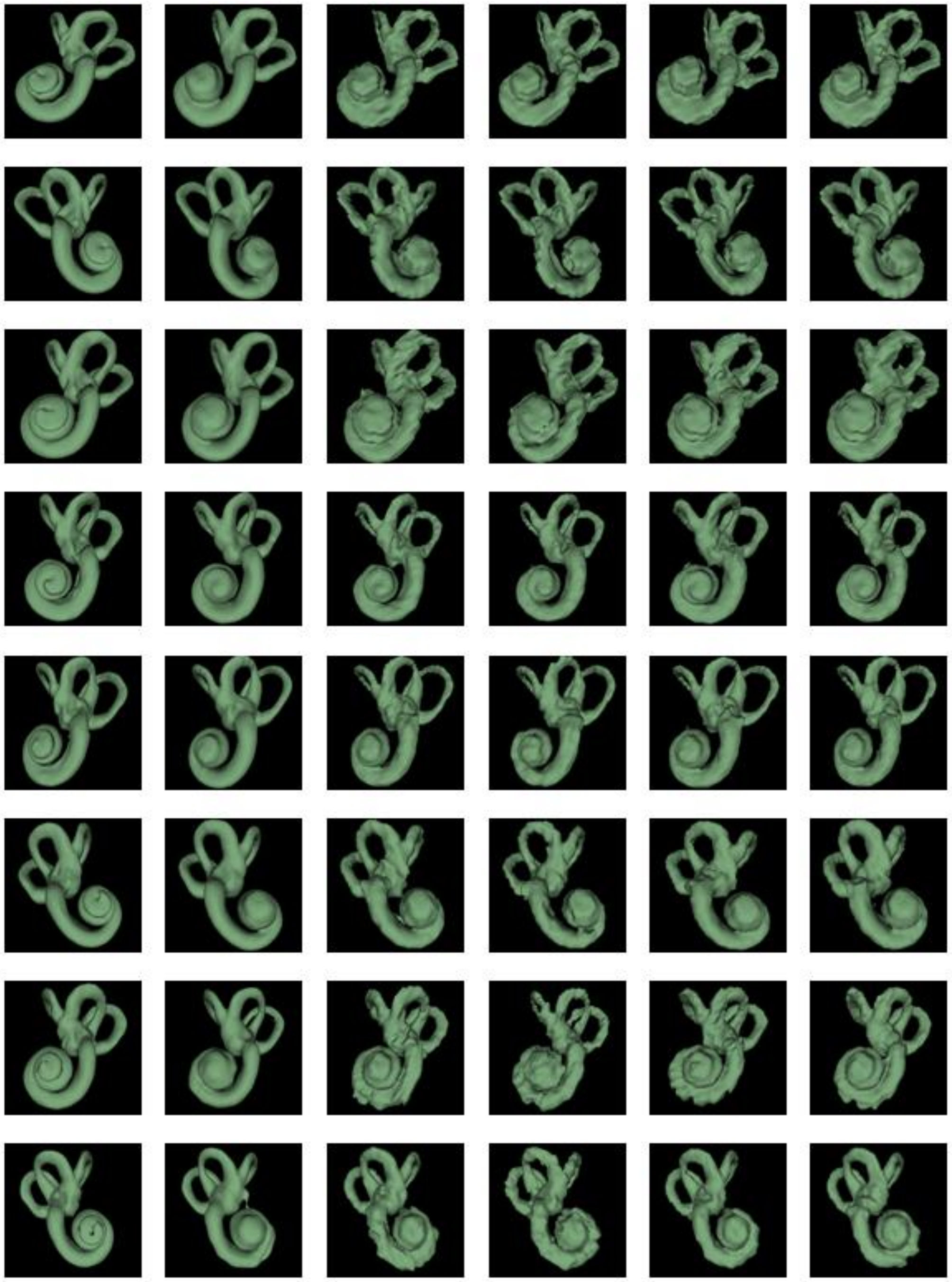


Figure 3.17: 3D renderings of ground truth segmentations, pipeline predictions, and expert segmentations. Each row corresponds to one of the eight cadaveric samples. Rows one to three correspond to 234×234×625um helical CT, rows four to six correspond to 300um isotropic CBCT, and rows seven and eight are 300um isotropic CBCT with implanted electrodes. The first column shows ground truth segmentations, the second column shows automated pipeline outputs, and columns three to six are expert segmentations.

3.3.2 Testing on Non-cadaveric Scans Without Electrodes

This section presents a qualitative and quantitative evaluation of the accuracy of the pipeline by comparing pipeline outputs to ground truth segmentations. The pipeline was tested using eight non-cadaveric clinical CBCT scans. The mean Dice score for the IE was 0.928 with a standard deviation of 0.013. The mean distance between the centroid of the RW predicted by the pipeline and the expert manually annotated centroid was 0.21mm with a standard deviation of 0.061mm. Figure 3.18 shows 3D renderings of pipeline predicted segmentations for all eight test set samples. Visually, there are no obvious IE segmentation errors such as bulbous false positives, or sections of the SCCs or cochlea missing. The predicted RW segmentations visually appear to have the anatomically correct saddle-like shape. Figure 3.19 demonstrates the relative scale of the average RW centroid error (0.21mm) compared to the anatomy. Figure 3.20 displays a pipeline RW segmentation which has a centroid distance error (0.205mm) closest to the mean of the eight samples (0.21mm).

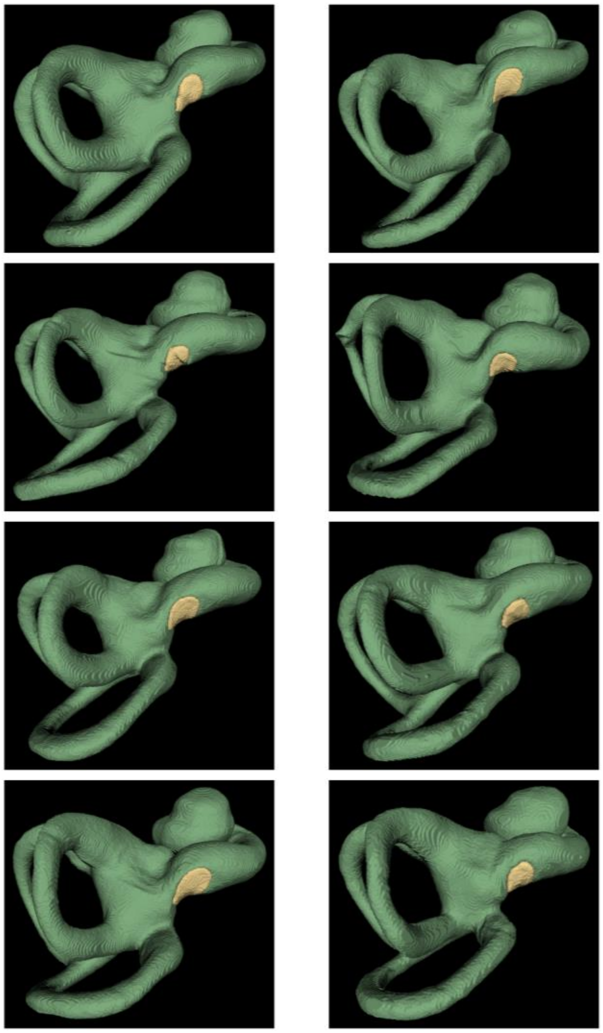


Figure 3.18: 3D visualizations of pipeline output segmentations for all eight non-cadaveric test set scans. The IE is depicted in green, and RW in yellow.

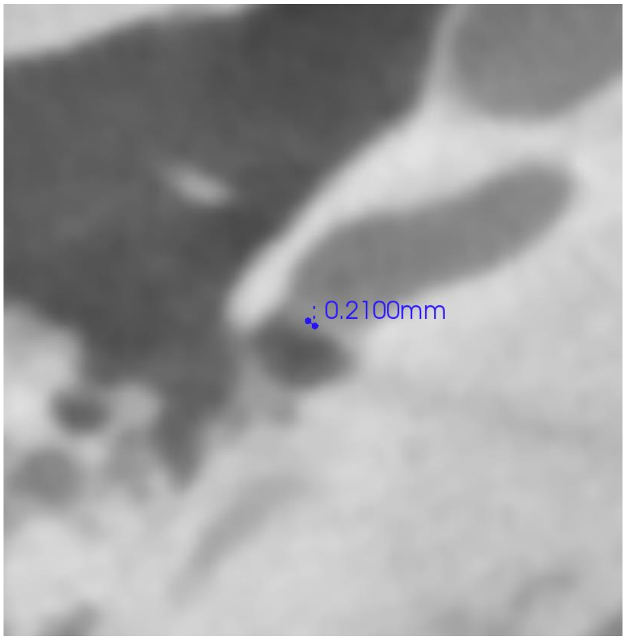


Figure 3.19: 2D visualization of the average RW segmentation centroid error.

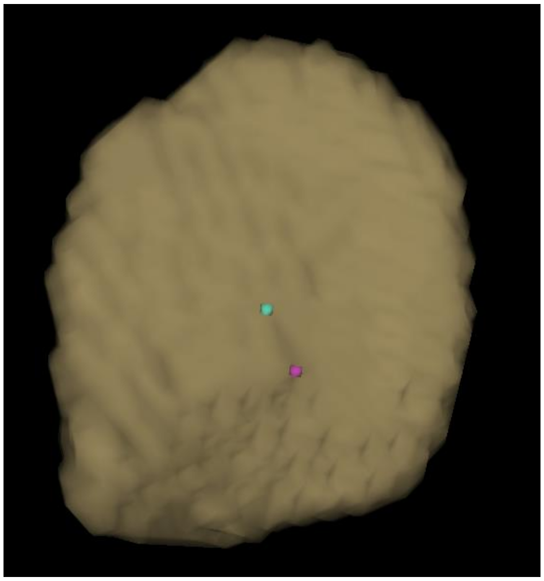


Figure 3.20: 3D visualization of a pipeline output RW segmentation (in translucent yellow) with the predicted centroid in cyan, and the expert labeled centroid in fuchsia.

3.4 Discussion

A pipeline was developed to automatically segment the IE and RW based on clinical CT scans with and without implanted electrodes. The approach achieves accurate automated segmentations rapidly, which facilitates implementation of the approach to a variety of settings including preoperative planning and surgical simulation. Adoption of the technique could help improve the safety of cochlear implantation, and hearing outcomes for CI recipients.

In the current state of the pipeline, input scans should be in clinical orientation and cropped to a region of interest extending approximately 5mm in all directions around the IE, as this was the standard for training, validation, and test data. Future work may explore the automation of these data preparation steps to fully automate the segmentation pipeline. This could be accomplished through the training of a separate network which segments the IE in complete temporal bone CT scans with lower resolution and a larger field of view. The output of this network could be utilized to automatically crop and orient scans prior to using the pipeline presented in this work.

To increase the accessibility of the model for clinical workflows while maintaining performance through high patch overlap, an inference server could be set up on a computer with performant hardware; this computer could accept scans from other computers, make predictions, and return the resultant segmentations. Alternatively, network performance should be evaluated using less patch overlap and more accessible hardware.

A limitation of the study was the small cadaveric test set size of eight samples. This limitation is a result of the difficulty associated with obtaining scans having both clinical and high-resolution acquisitions, especially SR-PCI. The nearest facility capable of SR-PCI quality images is located in the geographically distant Canadian Light Source facility in Saskatoon, and involves a competitive application for beam time. Secondly, the study involved comparison of pipeline segmentations to manual expert segmentations. Manual expert segmentations are time-consuming and busy surgeons and radiologists were unavailable to manually segment a large test set.

A robot-assisted CI procedure has recently been explored in its first clinical trial [25]. To automate the robot-assisted CI procedure, a high-quality segmentation of the IE and RW is necessary to determine safe drilling locations and an ideal electrode insertion angle. The present work presents an approach which could be used in an automated robot-assisted CI procedure. Surgically relevant surrounding structures such as the facial nerve should also be segmented to identify anatomical abnormalities and inform the robot-assisted approach. The facial nerve's anatomical course is highly variable, and the risk of facial paralysis resulting from nerve damage has led to intraoperative techniques which monitor the facial nerve [26]–[28]. Automated segmentation of the facial nerve would allow a robot-assisted approach to avoid facial nerve damage intraoperatively.

The automated pipeline was accurate in predicting the location of the RW centroid and apical tip of the cochlea; these results incentivize use of the pipeline in preoperative planning, electrode selection, and postoperative pitch-mapping. The exact cochlear duct length and angular length of the cochlea can be determined by tracing the lateral wall of the cochlea from RW to the apical tip [29]. Custom pitch mapping techniques which utilize known electrode angular depths result in reduced semitone errors when compared to one-size-fits-all pitch-mapping approaches [30]. To facilitate custom frequency mapping, postoperative electrode positions must be determined, where their angular depth measurements begin at the RW and end at the centroid of each electrode. If electrode positions are combined with the IE and RW segmentations presented in this work, sufficient information is available to automate postoperative customized pitch-mapping.

References

[1] P. Canzi *et al.*, "Anatomic variations of the round window niche: radiological study and related endoscopic anatomy," *Surg. Radiol. Anat. SRA*, vol. 41, no. 7, pp. 853–857, Jul. 2019, doi: 10.1007/s00276-019-02225-8.

[2] L. W. Helpard, S. A. Rohani, H. M. Ladak, and S. K. Agrawal, "Evaluation of Cochlear Duct Length Measurements From a 3D Analytical Cochlear Model Using Synchrotron Radiation Phase-Contrast Imaging," *Otol. Neurotol. Off. Publ. Am. Otol. Soc. Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol.*, vol. 41, no. 1, pp. e21–e27, Jan. 2020, doi: 10.1097/MAO.0000000000002420.

[3] J. c. Luers, K. b. Hüttenbrink, and D. Beutner, “Surgical anatomy of the round window—Implications for cochlear implantation,” *Clin. Otolaryngol.*, vol. 43, no. 2, pp. 417–424, 2018, doi: 10.1111/coa.13048.

[4] M. E. Timm *et al.*, “Patient specific selection of lateral wall cochlear implant electrodes based on anatomical indication ranges,” *PLoS ONE*, vol. 13, no. 10, p. e0206435, Oct. 2018, doi: 10.1371/journal.pone.0206435.

[5] M. W. Canfarotta *et al.*, “Relationship Between Electrocochleography, Angular Insertion Depth, and Cochlear Implant Speech Perception Outcomes,” *Ear Hear.*, vol. 42, no. 4, pp. 941–948, Aug. 2021, doi: 10.1097/AUD.0000000000000985.

[6] B. P. O’Connell *et al.*, “Insertion depth impacts speech perception and hearing preservation for lateral wall electrodes,” *The Laryngoscope*, vol. 127, no. 10, pp. 2352–2357, Oct. 2017, doi: 10.1002/lary.26467.

[7] O. Hilly *et al.*, “Depth of Cochlear Implant Array Within the Cochlea and Performance Outcome,” *Ann. Otol. Rhinol. Laryngol.*, vol. 125, no. 11, pp. 886–892, Nov. 2016, doi: 10.1177/0003489416660111.

[8] S.-C. Bae, Y.-R. Shin, and Y.-M. Chun, “Cochlear Implant Surgery Through Round Window Approach Is Always Possible,” *Ann. Otol. Rhinol. Laryngol.*, vol. 128, no. 6_suppl, pp. 38S-44S, Jun. 2019, doi: 10.1177/0003489419834311.

[9] T.-T. Lwin, A. Yoneyama, H. Maruyama, and T. Takeda, “Visualization Ability of Phase-Contrast Synchrotron-Based X-Ray Imaging Using an X-Ray Interferometer in Soft Tissue Tumors,” *Technol. Cancer Res. Treat.*, vol. 20, p. 15330338211010120, Apr. 2021, doi: 10.1177/15330338211010121.

[10] H. Li, N. Schart-Morén, S. A. Rohani, H. M. Ladak, H. Rask-Andersen, and S. Agrawal, “Synchrotron Radiation-Based Reconstruction of the Human Spiral Ganglion: Implications for Cochlear Implantation,” *Ear Hear.*, vol. 41, no. 1, pp. 173–181, Feb. 2020, doi: 10.1097/AUD.0000000000000738.

[11] M. Elfarnawany, S. R. Alam, S. a. Rohani, N. Zhu, S. k. Agrawal, and H. m. Ladak, “Micro-CT versus synchrotron radiation phase contrast imaging of human cochlea,” *J. Microsc.*, vol. 265, no. 3, pp. 349–357, 2017, doi: 10.1111/jmi.12507.

[12] J. E. Iglesias and M. R. Sabuncu, “Multi-Atlas Segmentation of Biomedical Images: A Survey,” *Med. Image Anal.*, vol. 24, no. 1, pp. 205–219, Aug. 2015, doi: 10.1016/j.media.2015.06.012.

[13] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, “3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support,” in *Intraoperative Imaging and Image-Guided Therapy*, F. A. Jolesz, Ed. New York, NY: Springer, 2014, pp. 277–

289. doi: 10.1007/978-1-4614-7657-3_19.

[14] G. Taubin, “Curve and surface smoothing without shrinkage,” in *Proceedings of IEEE International Conference on Computer Vision*, Jun. 1995, pp. 852–857. doi: 10.1109/ICCV.1995.466848.

[15] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation,” *Ieee Trans. Med. Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[16] L. A. Killewich, G. Falls, T. M. Mastracci, and K. R. Brown, “Factors affecting radiation injury,” *J. Vasc. Surg.*, vol. 53, no. 1, Supplement, pp. 9S-14S, Jan. 2011, doi: 10.1016/j.jvs.2010.07.025.

[17] R. Jaturapitpornchai, M. Matsuoka, N. Kanemoto, S. Kuzuoka, R. Ito, and R. Nakamura, “Newly Built Construction Detection in SAR Images Using Deep Learning,” *Remote Sens.*, vol. 11, p. 1444, Jun. 2019, doi: 10.3390/rs11121444.

[18] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers,” *Biomed. Opt. Express*, vol. 9, no. 7, pp. 3049–3066, Jul. 2018, doi: 10.1364/BOE.9.003049.

[19] R. Siddiqi, “Effectiveness of Transfer Learning and Fine Tuning in Automated Fruit Image Classification,” in *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, New York, NY, USA, Jul. 2019, pp. 91–100. doi: 10.1145/3342999.3343002.

[20] A. Myronenko, “3D MRI brain tumor segmentation using autoencoder regularization,” *ArXiv181011654 Cs Q-Bio*, Nov. 2018, Accessed: Dec. 17, 2021. [Online]. Available: http://arxiv.org/abs/1810.11654

[21] Y. Wu and K. He, “Group Normalization,” *ArXiv180308494 Cs*, Jun. 2018, Accessed: Dec. 07, 2021. [Online]. Available: http://arxiv.org/abs/1803.08494

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” *ArXiv170802002 Cs*, Feb. 2018, Accessed: Nov. 30, 2021. [Online]. Available: http://arxiv.org/abs/1708.02002

[23] S. Nikan *et al.*, “PWD-3DNet: A Deep Learning-Based Fully-Automated Segmentation of Multiple Structures on Temporal Bone CT Scans,” *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.*, vol. 30, pp. 739–753, 2021, doi: 10.1109/TIP.2020.3038363.

[24] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia,

"MeshLab: an Open-Source Mesh Processing Tool," p. 8.

[25] H. Daoudi *et al.*, "Robot-assisted Cochlear Implant Electrode Array Insertion in Adults: A Comparative Study With Manual Insertion," *Otol. Neurotol.*, vol. 42, no. 4, p. e438, Apr. 2021, doi: 10.1097/MAO.0000000000003002.

[26] L. P.-H. Li, J. K.-C. Chen, and D. H. Coelho, "Optimizing Location of Subdermal Recording Electrodes for Intraoperative Facial Nerve Monitoring," *The Laryngoscope*, vol. 131, no. 7, pp. E2329–E2334, Jul. 2021, doi: 10.1002/lary.29518.

[27] J. Ansó *et al.*, "Prospective Validation of Facial Nerve Monitoring to Prevent Nerve Damage During Robotic Drilling," *Front. Surg.*, vol. 6, p. 58, 2019, doi: 10.3389/fsurg.2019.00058.

[28] D. Dulak and I. A. Naqvi, "Neuroanatomy, Cranial Nerve 7 (Facial)," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2021. Accessed: Dec. 17, 2021. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK526119/

[29] R. W. Koch, H. M. Ladak, M. Elfarnawany, and S. K. Agrawal, "Measuring Cochlear Duct Length - a historical analysis of methods and results," *J. Otolaryngol. - Head Neck Surg. J. Oto-Rhino-Laryngol. Chir. Cervico-Faciale*, vol. 46, no. 1, p. 19, Mar. 2017, doi: 10.1186/s40463-017-0194-2.

[30] L. Helpard *et al.*, "An Approach for Individualized Cochlear Frequency Mapping Determined From 3D Synchrotron Radiation Phase-Contrast Imaging," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 12, pp. 3602–3611, Dec. 2021, doi: 10.1109/TBME.2021.3080116.

Chapter 4

A Deep Learning Solution for Automated Segmentation of the Inner Ear on Electrode-Implanted and Non-Implanted Computed Tomography with a Large Field of View and Varying Orientation

4.1 Introduction

Semi-automated segmentation approaches sometimes involve extensive manual data preparation. Common steps to prepare data for use of a CNN include cropping and rotation, which requires software and some technical expertise [1]. Manual steps in automated segmentation pipelines should be minimized whenever possible to facilitate their smooth integration into clinical workflows. Extensive manual steps necessitate personnel training and act as a barrier to entry for technical developments.

The goal of the present study was to quickly and accurately segment the IE in a variety of CT scans with minimal user input. This study presents an automated ML pipeline involving two patch-based 3D U-nets with residual connections. The novel pipeline is robust to rotation, can handle electrodes, and was found to result in accurate segmentations. Additionally, the segmentations produced by this work can be used to automatically preprocess scan volumes for utilization of the pipelines proposed in Chapter 3. This chapter builds on Chapter 3 by removing the constraint that data must be manually cropped and rotated before using the pipeline. This is the first known study to validate an automated IE segmentation approach on rotated CTs.

4.2 Materials and Methods

4.2.1 Image Datasets, Preparation, and Augmentation

As Chapter 4 builds on Chapter 3, some CTs were used in the development of pipelines presented in both chapters. Scans utilized in both chapters remained in the same data split (train, validation, test), which allowed for end-to-end testing of both pipelines. Cadaveric and non-cadaveric CTs with and without CIs were used to develop the automated segmentation pipeline. Table 4.1 summarizes all data used throughout the study. Scanner acquisition settings for non-cadaveric data were not available, and acquisition settings for cadaveric samples are the same as was presented in Chapter 3.

Table 4.1: Description of study data. ‘Sample Details’ lists data subsets which share image acquisition modalities. Each bullet point in ‘Input Acquisitions’ represents one acquisition and specifies if an electrode was implanted in the corresponding scans. Acquisitions used in the test set appear in bold. ‘Ground Truth Modality’ and ‘Ground Truth Segmentation Method’ describe the imaging modality and method used to form ground truth segmentations in the respective data subset. The number of samples assigned to train, validation, and test sets are specified in ‘Data Split’.

Sample Details	Input Acquisitions	Ground Truth Modality	Ground Truth Segmentation Method	Data Split
64 cadaveric	● CBCT 300um isotropic ● Helical CT 488um×488um×625 um Soft Tissue Protocol ● Helical CT 488um×488um×625 um Temporal Bone protocol ● Helical CT 488um×488um×1250um Soft Tissue Protocol ● Helical CT 488um×488um×1250um Temporal Bone	Micro-CT 50um isotropic	IE: Micro-CT 50um isotropic manually segmented.	Train: 46 Validation: 18

	protocol			
84 non-cadaveric	**CBCT variety of clinical resolutions**	CBCT	IE: Preliminary network inference and manual correction. IE (test set samples): Manual segmentation of scan upsampled to 154um isotropic.	Train: 53 Validation: 23 Test: 8
32 cadaveric	Helical CT 234×234×625um	Micro-CT 154um isotropic	IE: Preliminary network inference and manual correction	Train: 23 Validation: 9
12 cadaveric	● **CBCT 300um isotropic, electrode-implanted** ● **Helical CT 625×234×234um, electrode-implanted** ● Helical CT 625×488×488um Soft Tissue Protocol, electrode-implanted ● Helical CT 625×488×488um Temporal Bone protocol, electrode-implanted ● Helical CT 1.25×488×488um Soft Tissue Protocol electrode-implanted ● Helical CT 1.25×488×488um Temporal Bone protocol, electrode-implanted	Micro-CT 154um isotropic	IE: Preliminary network inference and manual correction IE (test set samples): Micro-CT 154um isotropic upsampled to 50um isotropic and manually segmented	Train: 8 Validation: 2 Test: 2

6 non-cadaveric	CBCT 200um isotropic, electrode-implanted	Qualitatively best pre-operative CBCT available	IE: Preliminary network inference and manual correction	Train: 4 Validation: 2
7 non-cadaveric	CBCT variety of clinical resolutions, electrode-implanted	Qualitatively best pre-operative CBCT available	IE: Preliminary network inference and manual correction	Train: 5 Validation: 2

Image acquisitions for each sample were rigidly registered. All scans and segmentations included in the study, regardless of original voxel spacing, were resampled to 154um isotropic voxel spacing to ensure a consistent input resolution for CNN training. A larger voxel size was selected than in Chapter 3 to compensate for the greatly increased scan field of view in the present work.

Two graduate students manually segmented the IE in 64 50um isotropic voxel spacing micro-CT scans. To create ground truth segmentations for all other data which had no high-resolution acquisition, a multi-step process was implemented. First, scans were manually cropped to a 5mm region surrounding the IE and upsampled to 50um isotropic voxel spacing. The pipelines presented in Chapter 3 were used to generate preliminary IE segmentations. Each segmentation was then manually checked by a graduate student and segmentation imperfections were corrected manually.

Image augmentations were applied to all scans to expand the dataset size and add variance. All scans used in network training were manually cropped to a spherical region with a diameter of 100mm surrounding the IE (Figure 4.1). All voxels outside this spherical region were set to the lowest voxel intensity (black) in their respective images. Spherical cropping was necessary to limit the size of scan volumes. This form of cropping requires less expertise and time than the approach utilized to prepare data in Chapter 3, as spherical cropping can be achieved based on a single user click which represents the center of the spherical crop. All scans were flipped in the sagittal plane and modified with two levels of additive Gaussian noise as was described in Chapter 3. Scans of all resolutions were resampled to 154um isotropic voxel spacing with B-spline interpolation, and nearest neighbor resampling was used to resample segmentations to 154um isotropic voxel spacing. Scans and segmentations were resampled to the same image grid to ensure alignment. All scans were also randomly rotated from -180 to 180 degrees in each of the x, y, and z axis; this rotation functionally achieved full 360-degree coverage such that all rotational variations were possible. The pipeline was designed to perform best on data in clinical orientation, as half of network training data was not rotated and the other half was equally distributed random rotations. In most cases, scans input to the network should be in clinical orientation, unless cadaveric samples were acquired without careful specimen positioning, or if CT metadata which tracks rotation was lost.

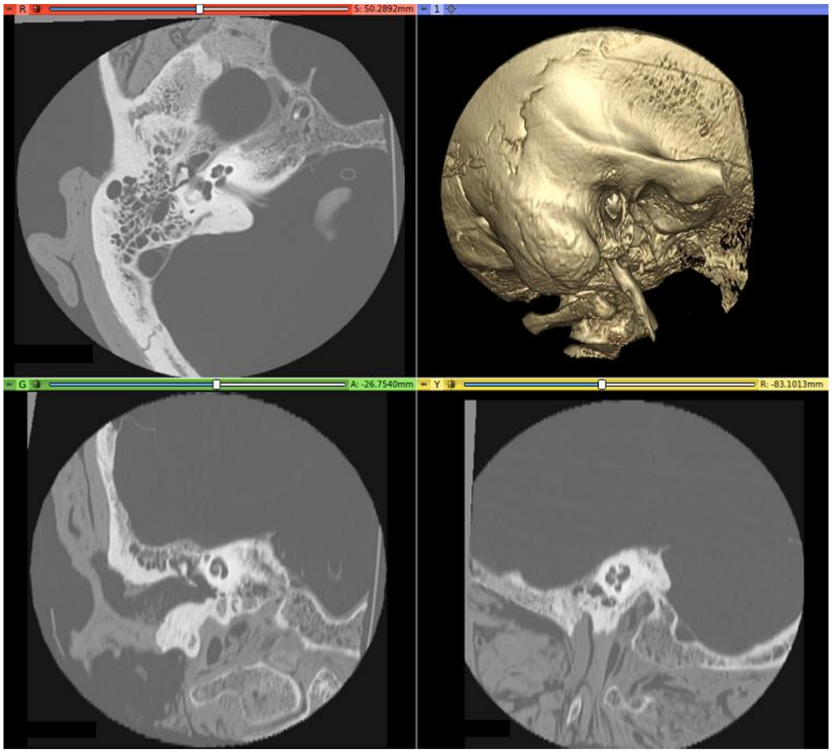


Figure 4.1: Cadaveric helical CT study sample cropped to a 100mm diameter sphere surrounding the IE. From top to bottom and left to right: axial view, 3D rendering of bone-density scan voxels, coronal view, sagittal view. 2D and 3D visualization was done with 3DSlicer (v4.11, https://www.slicer.org/).

4.2.2 Segmentation Pipeline

Of the 180 samples included in the no-electrode network, 122 were assigned to the training set, 50 to the validation set, and eight to the test set. Of the 25 samples included in the electrode-implanted network, 17 were assigned to the training set, six to the validation set, and two to the test set. All image acquisitions of the same sample were assigned to the same set. The networks were trained using the same workstation and software setup as was described in Chapter 3.

The no-electrode network was trained for a total of 23 epochs. The electrode-implanted network was trained for 25 epochs with initial weights set to values learned during the no-electrode network training. Because the dataset had far more scans without electrodes than with, the 23 epochs of the no electrode training involved far more iterations than the 25 epochs of the training associated with the electrode-implanted network.

The network architecture utilized was similar to that which was presented in Chapter 3 with the exception that the final activation function was sigmoid, as is appropriate for binary classification problems. The output of the network, after being passed through the sigmoid function, was a two-channel probability map with values indicating the model's certainty that each voxel belonged to class 0 (background), or 1 (IE). The patch size was 288×256×256 with a batch size of one. The ADAM learning rate optimizer was used with a learning rate of 0.0001. Voxel intensities were normalized to zero mean and unit standard deviation, then shifted during training iterations between -30% and +30%. The network which achieved the maximum validation Dice score on the IE label was selected as the best network.

The loss function was the same focal loss $(-(1-p_t)^{\gamma} log(p_t), \gamma = 2)$ and Dice loss $(1 - \frac{2|X \cap Y|}{(|X|+|Y|)})$ hybrid as was described in Chapter 3. Loss associated with the background class was not included in the loss calculation, as inclusion of background loss contributions allow the loss function to be dominated by the background class. As a result of the increased scan field of view, the vast majority (>99%) of scan voxels correspond to the background class. Balanced patch sampling was used in which each training and validation iteration involved a 50% chance of image patches being centered around an IE or background voxel. This patch sampling approach ensured the network would see sufficient examples including the IE, rather than being dominated by full patches of background. Inference utilized a sliding window approach with 70% patch overlap. Network outputs were post-processed to complete the segmentation pipeline. First, the argmax function was used to convert the two-channel probability map produced by the final sigmoid activation function into a finite segmentation label map with values of 0 or 1. Island removal was then applied to the IE segmentation which ensured only IE islands composed of more than 4,000 voxels were kept; the threshold of 4,000 voxels was selected experimentally by measuring the size of false positives produced by the network. This threshold was used rather than keeping only the largest island to handle cases where false negatives (IE voxels incorrectly classified as background) could separate portions of the IE. Smoothing of segmentation volumes was deemed unnecessary as patch lines were minimal as a result of the increased 154um voxel spacing.

Metrics used to evaluate the segmentation pipeline include: Dice coefficient (DC) ($\frac{2|X \cap Y|}{(|X|+|Y|)}$), Maximum Hausdorff Distance (MHD) ($max\{sup_{x \in X}(inf_{y \in Y} d(x,y)), sup_{y \in Y}(inf_{x \in X} d(x,y))\}$), and Average Hausdorff Distance (AHD).

4.3 Results

Network inference was tested on the same alternate workstation as was described in Chapter 3. For a scan with a physical dimensionality of 75mm×90mm×82mm, the pipeline took 175 s (upsampling, 1 s; network inference, 135 s; and post-processing, 39 s) to segment the IE with 70% patch overlap.

4.3.1 Testing on Cadaveric Scans with Electrodes

This section presents a qualitative and quantitative evaluation of the accuracy of the pipeline by comparing pipeline outputs to ground truth segmentations. The pipeline was tested using two cadaveric samples with implanted electrodes which have both CBCT and helical CT acquisitions. Metrics are also presented for the four scans after random rotational augmentations were applied. Quantitative results are shown in Table 4.2. Results followed an expected trend where the pipeline performed better on scans in standard clinical orientation than randomly rotated; half of the data used in network training and validation was in clinical orientation, meaning the CNN saw fewer examples of data in all other rotational variants. The pipeline also performed better on CBCT than on helical CT, as expected due to CBCT acquisitions having superior imaging characteristics compared to helical CT. Pipeline performance can be seen visually (Figure 4.2) where rotated samples have far more false positives compared to the clinical orientation counterparts.

Table 4.2: Mean values are presented for all metrics. Data ranges are given in parenthesis. Metrics include Inner Ear Dice Coefficient (IE DC), Inner Ear Average Hausdorff Distance (IE AHD), and Inner Ear Maximum Hausdorff Distance (IE MHD). Metrics are presented for four test set subsets.

CT Modality	IE DC	IE AHD (mm)	IE MHD (mm)
CBCT 300um isotropic electrode-implanted	0.859 (0.002)	0.116 (0.005)	0.691 (0.064)
CBCT 300um isotropic rotated electrode-implanted	0.837 (0.018)	0.163 (0.008)	1.084 (0.044)
Helical CT 234×234×625um electrode-implanted	0.838 (0.006)	0.140 (0.006)	0.780 (0.183)
Helical CT 234×234×625um rotated electrode-implanted	0.795 (0.038)	0.186 (0.020)	1.091 (0.050)

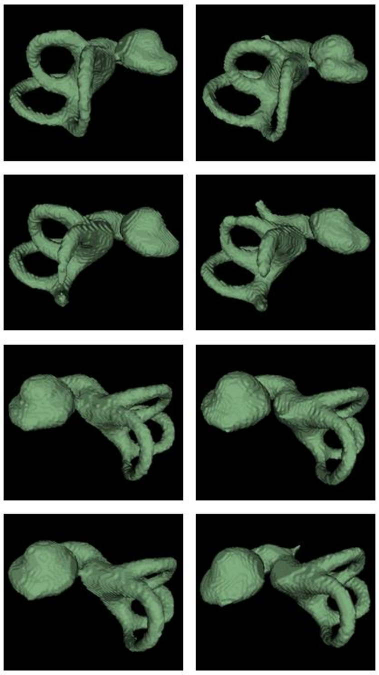


Figure 4.2: 3D renderings of pipeline output segmentations for all implanted test set scans. Pairs of side-by-side images correspond to a single sample; the right-hand image of each pair corresponds to the randomly rotated sample.

4.3.2 Testing on Non-cadaveric Scans Without Electrodes

This section presents a qualitative and quantitative evaluation of the accuracy of the pipeline by comparing pipeline outputs to ground truth segmentations. The pipeline was tested using eight non-cadaveric clinical CBCT scans. Metrics are also presented for the eight scans after random rotational augmentations were applied. Table 4.3 depicts quantitative results. Results followed an expected trend where the pipeline performed better on scans in standard clinical orientation than randomly rotated. Based on 3D renderings of pipeline predicted segmentations it is apparent that the rotated samples have more imperfections (Figure 4.3).

Table 4.3: Quantitative results with std dev including Inner Ear Dice Coefficient (IE DC), Inner Ear Average Hausdorff Distance (IE AHD), Inner Ear Maximum Hausdorff Distance (IE MHD).

CT Modality	IE DC	IE AHD (mm)	IE MHD (mm)
Clinical CBCT	0.895 ± 0.019	0.101 ± 0.021	1.07 ± 0.109
Clinical CBCT rotated	0.885 ± 0.017	0.107 ± 0.023	1.07 ± 0.125

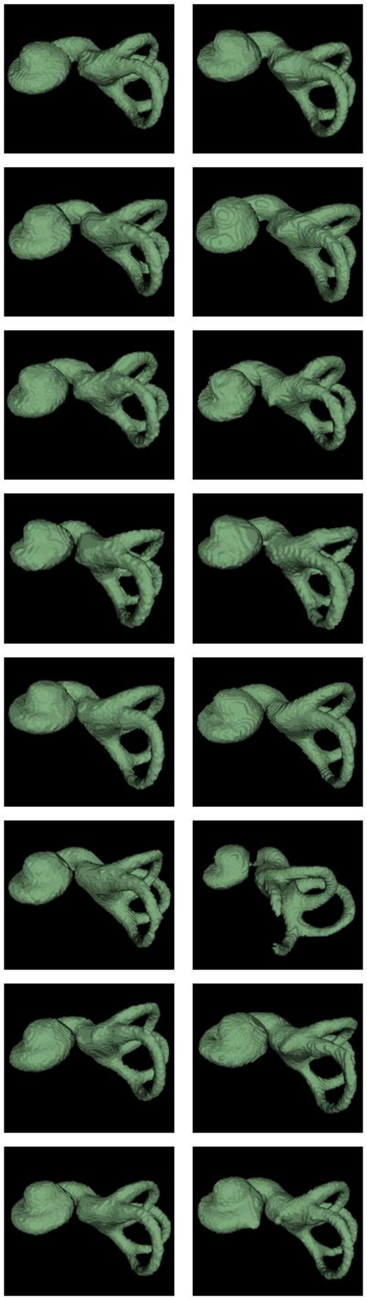


Figure 4.3: 3D renderings of predicted segmentations for non-implanted test set scans. Pairs of side-by-side segmentations correspond to a single sample; the right-hand segmentation of each pair corresponds to the randomly rotated sample.

4.4 Discussion

A pipeline was developed to automatically segment the IE on clinical CT scans in any orientation, with and without implanted electrodes. The approach is robust and requires minimal preprocessing steps. The approach quickly and accurately segments the IE in a variety of CT scans. A scan volume can be cropped based on the IE segmentation predicted by this pipeline to automate the data preprocessing for use of the pipelines presented in Chapter 3. This network is useful as a standalone to identify significant abnormalities in the shape of IE structures such as the cochlea, SCCs, and vestibule; a higher resolution segmentation, as is produced by pipelines in Chapter 3, may be unnecessary to identify gross anatomical abnormalities. Additionally, the 154um segmentations produced by the pipelines presented in Chapter 4 may be sufficient for some surgical simulation platforms which have trouble rendering the amount of data present in a higher resolution segmentation volume; this may arise in cases where the system running a surgical simulation is limited in computational power.

Rotational augmentations were necessary in the present work to alleviate the need for manual preprocessing steps. An alternative approach to achieve segmentation of rotated scans could be training CNNs from the Chapter 3 pipelines to be performant on rotated samples. This idea was explored through the training of two CNNs with an architecture similar to that of Chapter 3, but with only a single output channel, meaning the RW was not segmented. One network was trained on a dataset twice as large, where each sample was augmented with random rotations, as was described in Chapter 4. The other network was trained with only the clinical orientation data. Figure 4.4 displays a comparison of the pipeline performance on two test set clinical orientation scans.

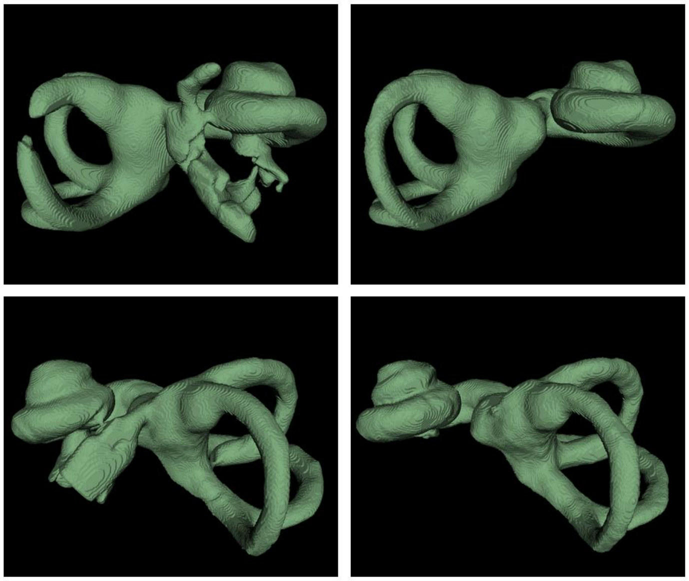


Figure 4.4: 3D renderings of pipeline segmentations with and without rotational augmentations. The left and right columns show pipeline outputs with and without rotational augmentations respectively. The top row is based on cadaveric 300um isotropic CBCT, and the second row non-cadaveric CBCT.

As the networks were tested on data in clinical orientation, poor performance of the network trained with rotational augmentations indicates that as the model learned to perform on rotated scans, there was a performance trade-off for clinical orientation data. The most likely explanation for the poor performance of the 50um isotropic spacing network with rotated samples is the limited spatial field of view of the image patches. Although the patch size is 288×256×256 voxels in all trained networks, this corresponds to a 14.4mm×12.8mm×12.8mm region in 50um isotropic scans, and a 44.4mm×39.4mm×39.4mm region in 154um isotropic scans. The increased amount of spatial information present in patches of the 154um network provided sufficient information to recognize the IE in rotated positions, whereas the 50um network was unable to do so. Models with larger patch sizes have been found to have improved performance in computer vision tasks [2].

The automated pipeline presented in this chapter accurately predicts the IE in a variety of CT scans rotated in any orientation. It is acceptable that IE segmentations produced by the model described in this chapter are less visually appealing than predictions from the model presented in Chapter 3; this is because the primary use for this model is to automatically preprocess scans such that the Chapter 3 model can be utilized with minimal user input.

References

[1] O. Diaz *et al.*, “Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools,” *Phys. Med.*, vol. 83, pp. 25–37, Mar. 2021, doi: 10.1016/j.ejmp.2021.02.007.

[2] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers,” *Biomed. Opt. Express*, vol. 9, no. 7, pp. 3049–3066, Jul. 2018, doi: 10.1364/BOE.9.003049.

Chapter 5

Conclusion

5.1. Summary

ML is quickly transforming the field of healthcare by providing assistance to doctors in an ever-growing number of ways. Contrary to the common “Will ML replace doctors?” narrative, a more important question to be asked is “How can ML help doctors and improve healthcare?”. The presented work demonstrates ML assisting doctors to improve healthcare, as automated segmentations of the IE and RW are useful tools for doctors, but too time consuming to be done manually in clinical practice.

Within this thesis, Chapter 3 demonstrated that CNNs can quickly and accurately generate detailed segmentations of the IE and RW. The method proved effective in electrode-implanted and non-implanted CT scans on a variety of CTs including cadaveric and non-cadaveric scans, as well as CBCT and helical CT. Tightly cropped CT scans in a standard clinical orientation were used to train the network. Use of this model as a stand-alone would require manual processing to ensure that scans are in clinical orientation and cropped to a region around the IE.

Chapter 4 presented CNNs which segment the IE in electrode-implanted and non-implanted scans on larger field of view CTs which may be rotated in any spatial orientation. Scans used in this network were cropped to a 100mm diameter sphere surrounding the IE, and were rotated randomly in 360 degrees. This network can be used to achieve an IE segmentation in any CT scan. When the Chapter 4 pipelines are used in conjunction with the pipelines presented in Chapter 3, automatic cropping of scan volumes can be done to generate detailed IE and RW segmentations with minimal preprocessing.

5.2. Future Work

To improve the utility of the automated segmentation approach, nearby structures including the ossicles, facial nerve, and sigmoid sinus could be added to the network predictions. Expanding the number of structures segmented allows the pipeline to be applicable to otolaryngological procedures which involve numerous structures of the skull. Automated segmentation of these structures would allow for surgical planning, training, and robot assisted approaches to be further developed for many otolaryngological surgeries.

A common concern with the application of CNNs to medical imaging problems is a lack of robustness to a wide variety of imaging data. To create a robust model, CNNs should be trained with medical scans from many institutions worldwide from all manufacturers and with all combinations of acquisition settings. To alleviate this problem in deep learning models, federated learning has been introduced. Federated learning facilitates easy and ethical sharing of data amongst institutions for the use of ML model training [1]. To minimize ethical concerns related to data sharing, federated learning allows multiple institutions to contribute data which trains a centralized model, but keeps data invisible to each institution. Federated learning could be applied to continue the training of networks deployed in this work, which should result in more robust models.

Clinical trials are currently underway which aim to determine the effectiveness of customized CI frequency-maps; these frequency maps are based on manually annotated segmentations and measurements. To evaluate the effects of the automated steps presented in this work, quantitative analysis can be done which compares measurements taken on the manually segmented structures to that of the automated pipeline. Differences in electrode frequencies between the manual and automated approaches can be calculated directly, and a further clinical trial can find clinical implications of differences between the approaches.

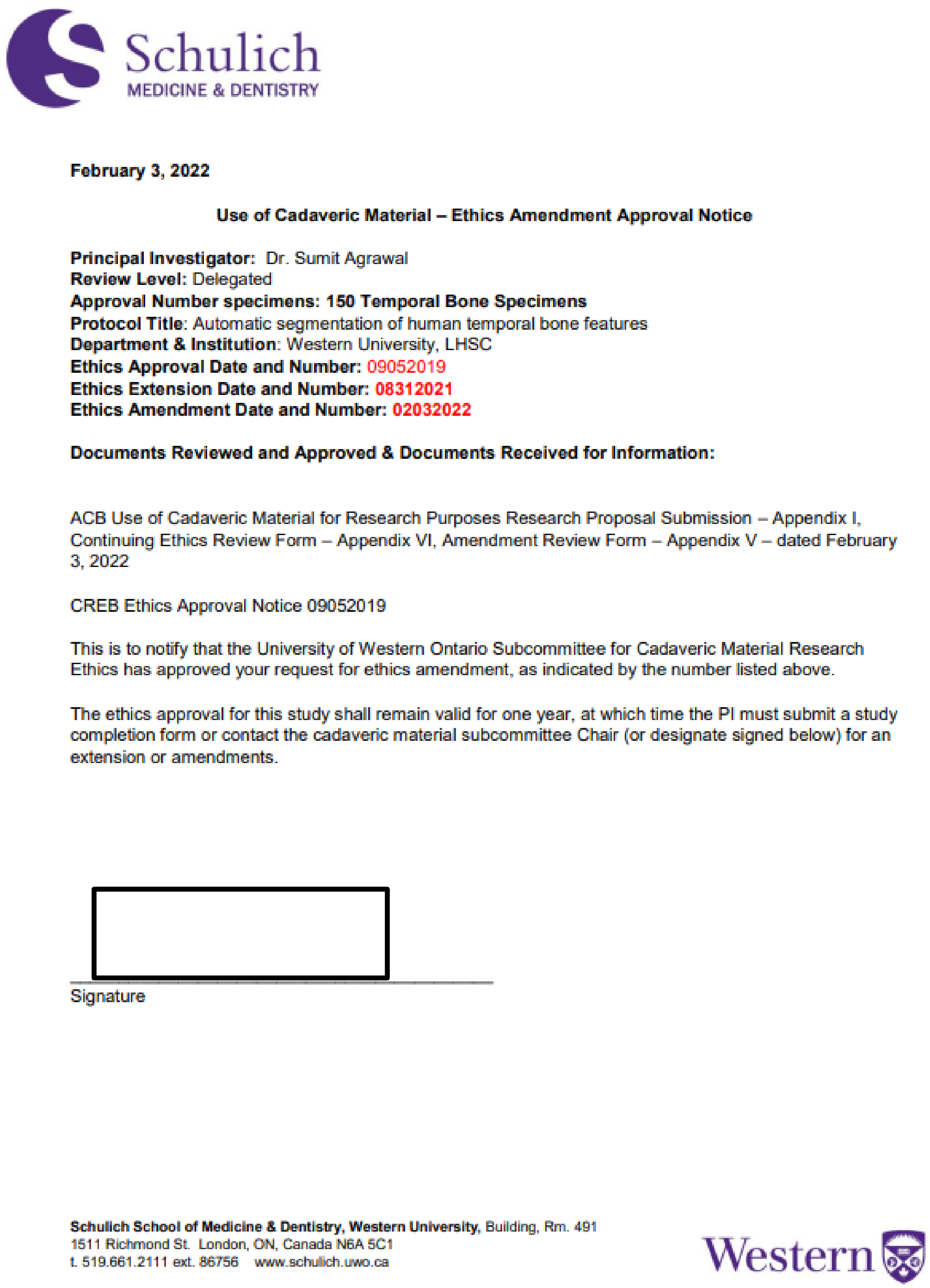
Lastly, software integration steps should be completed to combine the automated pipelines presented in this work to surgical training and planning software. This integration facilitates improved surgical planning and training, as 3D volumes of patient anatomy could be visualized for each clinical patient.

References

[1] M. J. Sheller *et al.*, “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data,” *Sci. Rep.*, vol. 10, no. 1, p. 12598, Jul. 2020, doi: 10.1038/s41598-020-69250-1.

Appendices

Appendix A: Cadaveric Ethics Approval



Schulich
MEDICINE & DENTISTRY

February 3, 2022

Use of Cadaveric Material – Ethics Amendment Approval Notice

Principal Investigator: Dr. Sumit Agrawal
Review Level: Delegated
Approval Number specimens: 150 Temporal Bone Specimens
Protocol Title: Automatic segmentation of human temporal bone features
Department & Institution: Western University, LHSC
Ethics Approval Date and Number: 09052019
Ethics Extension Date and Number: **08312021**
Ethics Amendment Date and Number: **02032022**

Documents Reviewed and Approved & Documents Received for Information:

ACB Use of Cadaveric Material for Research Purposes Research Proposal Submission – Appendix I, Continuing Ethics Review Form – Appendix VI, Amendment Review Form – Appendix V – dated February 3, 2022

CREB Ethics Approval Notice 09052019

This is to notify that the University of Western Ontario Subcommittee for Cadaveric Material Research Ethics has approved your request for ethics amendment, as indicated by the number listed above.

The ethics approval for this study shall remain valid for one year, at which time the PI must submit a study completion form or contact the cadaveric material subcommittee Chair (or designate signed below) for an extension or amendments.

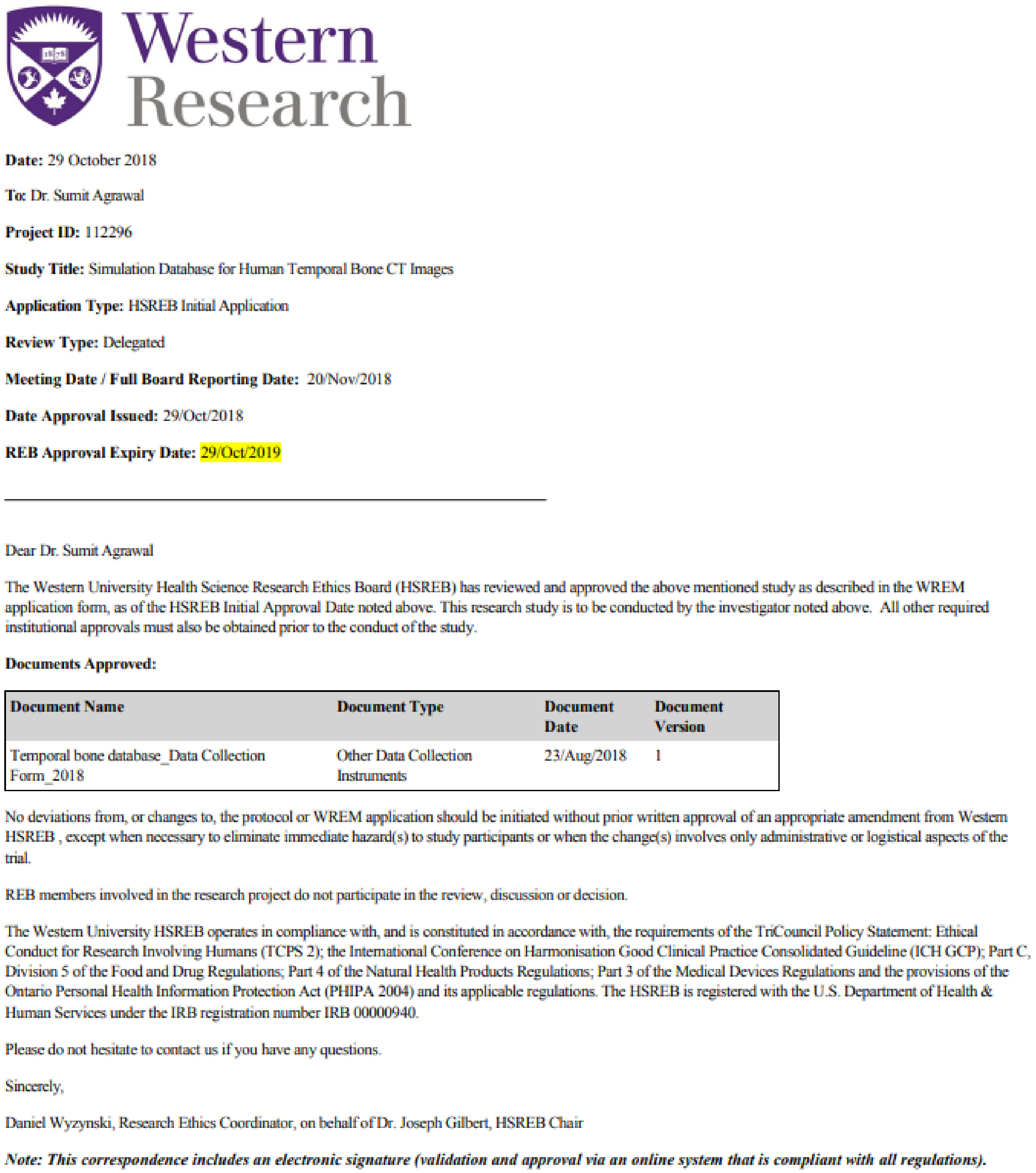
[redacted]

Signature

Schulich School of Medicine & Dentistry, Western University, Building, Rm. 491
1511 Richmond St. London, ON, Canada N6A 5C1
t. 519.661.2111 ext. 86756 www.schulich.uwo.ca

Western

Appendix B: Clinical Ethics Approval



Western
Research

Date: 29 October 2018

To: Dr. Sumit Agrawal

Project ID: 112296

Study Title: Simulation Database for Human Temporal Bone CT Images

Application Type: HSREB Initial Application

Review Type: Delegated

Meeting Date / Full Board Reporting Date: 20/Nov/2018

Date Approval Issued: 29/Oct/2018

REB Approval Expiry Date: 29/Oct/2019

Dear Dr. Sumit Agrawal

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above mentioned study as described in the WREM application form, as of the HSREB Initial Approval Date noted above. This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Temporal bone database_Data Collection Form_2018	Other Data Collection Instruments	23/Aug/2018	1

No deviations from, or changes to, the protocol or WREM application should be initiated without prior written approval of an appropriate amendment from Western HSREB , except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University HSREB operates in compliance with, and is constituted in accordance with, the requirements of the TriCouncil Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Daniel Wyzynski, Research Ethics Coordinator, on behalf of Dr. Joseph Gilbert, HSREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Curriculum Vitae

Name:

Kyle Rioux

Post-Secondary Education and Degrees:

2020-2022
Master of Engineering Science, Electrical and Computer Engineering, Software
Western University
London, Ontario, Canada

2016-2020
Bachelor of Engineering Science, Software Engineering
Western University
London, Ontario, Canada

Honours and Awards:

2019 NSERC Undergraduate Student Research Award
2016-2020 Dean’s Honour List
2019-2022 Academic All-Canadian

Related Work Experience:

2020-2021 Graduate Teaching Assistant (Information Security - SE 4472/ECE 9064)

Publications and Presentations:

K. A. Rioux, L. W. Helpard, H. M. Ladak, S. K. Agrawal, “A deep learning solution for automated segmentation of the inner ear and round window on electrode-implanted and non-implanted clinical computed tomography” *Scientific Reports*. Submitted.