Deep Learning Methods in 3D Computed Tomography Images for Implantable Devices

Lopez Diez, Paula

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
Doctor of Philosophy

DTU Compute
Department of Applied Mathematics and Computer Science

Deep Learning Methods in 3D Computed Tomography Images for Implantable Devices

Paula López Diez

Kongens Lyngby 2024

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Matematiktorvet
Building 303B
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary

Implantable medical devices have revolutionized modern medicine, offering life-saving interventions and significantly improving the quality of life for many patients. The process of implanting a medical device involves several key stages to ensure patient safety and achieve optimal results. These personalized stages include medical evaluation, diagnosis, therapy selection, surgical planning, surgical procedure, and postoperative follow-up.

This thesis focuses on developing advanced 3D image analysis methods to enhance preoperative tasks in CT images involving procedures for medical implantable devices. More specifically for cochlear implantation and left atrial appendage occlusion, these procedures involve navigating complex and variable anatomical structures, where accurate diagnosis and planning are crucial for successful outcomes.

A significant part of the work is dedicated to cochlear implantation, where the intricate anatomy of the inner ear presents substantial challenges. The thesis introduces deep reinforcement learning techniques to detect and characterize anatomical landmarks in the inner ear, which are vital for guiding CI procedures. These methods enable precise identification of neural structures and head orientation in CT scans, improving the safety and effectiveness of these interventions. We further extended the deep reinforcement approach to detect abnormal anatomies of the inner ear, which was both pioneering work for this anatomical region and also a step forward for anomaly detection in 3D medical images.

Another important contribution of this thesis is the development of the first automated approach for classifying congenital inner ear anomalies. This unsupervised method utilizes latent space representations of 3D point clouds to categorize the different types of congenital malformations, providing latent space representation that is representative of these complex and rare inner ear pathologies. This approach addresses a critical gap in the field, offering support for the diagnosis of these difficult conditions. The thesis also extends its focus to left atrial appendage occlusion procedures, using a graph-based model to analyze 3D cardiac structures. These models are designed to capture key anatomical features that influence procedural outcomes, helping clinicians navigate the complex and variable cardiac anatomy more effectively. We focus on exploiting the explainability features of this method to raise new clinical questions.

Overall, this thesis makes substantial contributions to the field of medical imaging by integrating cutting-edge deep learning techniques with specific clinical practice objectives. The research, not only advances the technical capabilities of medical image analysis but highlights the importance of collaboration between computer scientists and clinical experts to translate these innovations to enhance patient care.

Resume

Implanterbare medicinske enheder har revolutioneret moderne medicin, ved at tilbyde livsreddende interventioner og derved markant forbedre livskvaliteten for mange patienter. I processen med at implantere en medicinsk enhed er der er flere nøglefaser med det formål at sikre patientens sikkerhed og opnå optimale resultater. Disse individuelle faser inkluderer medicinsk evaluering, diagnose, valg af terapi, kirurgisk planlægning, kirurgisk procedure og postoperativ opfølgning.

Denne afhandling fokuserer på udvikling af avancerede 3D-billedanalysemetoder med det formål at forbedre præoperative opgaver på CT-billeder, som involverer procedurer for medicinske implanterbare enheder. Mere specifikt for cochlear implantation (CI) og lukning af auriklen (LAAO-procedure). Disse procedurer involverer navigation i komplekse og variable anatomiske strukturer, hvor nøjagtig diagnose og planlægning er afgørende for et vellykket resultat.

En væsentlig del af arbejdet er dedikeret til cochlear implantation, hvor det indre øres komplekse anatomi udgør en betydelig udfordring. Afhandlingen introducerer teknikker inden for deep reinforcement learning til at detektere og karakterisere anatomiske landemærker i det indre øre, som er afgørende for at udføre CI-operationer. Disse metoder muliggør præcis identifikation af nerve-strukturer og hovedets orientering i CT-scanninger, hvilket forbedrer sikkerheden og effektiviteten af disse interventioner. Vi udvidede yderligere metoden til deep reinforcement learning til at detektere unormal anatomi i det indre øre. Dette er innovativt arbejde inden for dette anatomiske område og et fremskridt for detektion af uregelmæssigheder i 3D billeder til medicinsk brug.

Et andet vigtigt bidrag fra denne afhandling er udviklingen af den første automatiserede metode til klassificering af medfødte anomalier i det indre øre. Denne usuperviserede metode udnytter repræsentationer af 3D-punktskyer i et latent rum til at kategorisere de forskellige typer af medfødte misdannelser. Repræsentation i et latent rum er repræsentativ for disse komplekse og sjældne patologier i det indre øre. Denne tilgang adresserer en kritisk mangel i feltet og tilbyder støtte til diagnosen af disse vanskelige tilstande.

Afhandlingen udvider også sit fokus til LAAO-procedurer ved brug af grafbaserede modeller til at analysere 3D-hjertestrukturer. Disse modeller er designet til at fange centrale anatomiske træk, der påvirker de operationsmæssige resultater, hvilket hjælper klinikere med mere effektivt at navigere i den komplekse og varierede hjerteanatomi. Vi fokuserer på at udnytte funktionerne i disse metoder der understøtter forklarlighed for at rejse kliniske spørgsmål.

Samlet set yder denne afhandling betydelige bidrag til området for medicinsk billedanalyse ved at integrere avancerede teknikker fra deep learning med klinisk praksis. Forskningen fremmer ikke kun de tekniske kapaciteter inden for medicinsk billedanalyse, men understreger også vigtigheden af samarbejde mellem specialister inden for avanceret modellering og kliniske eksperter for at oversætte disse innovationer til forbedret patientpleje.

Preface

This Ph.D. thesis was prepared in the Visual Computing section of the DTU Compute department at the Technical University of Denmark (DTU) in fulfillment of the requirements for acquiring a Ph.D. degree in medical image analysis. The work presented in this thesis is funded by William Demant Fonden. The project was carried out in collaboration with Oticon Medical, the Tashkent International Clinic, and Rigshospitalet. The research was conducted for three years, from September 2021 to August 2024.

The topic of this PhD is deep learning methods applied to 3D CT images for medical device therapy planning. Ten peer-reviewed papers accepted for publication were written during this thesis. Seven of those are included in this thesis as main contributions and constitute the principal research output of this project. Furthermore, several diverse student projects have been supervised during this time. An overview of all these contributions can be seen in the list of contributions section on page v.

The main contributions are covered in seven peer-reviewed papers, six in conference proceedings (*A*, *B*, *C*, *D*, *F* and *G*) and a journal paper (*E*). These seven papers are available in Appendix A-G as a part of the thesis.

The project was supervised by Rasmus Reinhold Paulsen from DTU Compute, Jan Margeta from Oticon Medical and KardioMe, and François Patou from Oticon Medical.

The main research activities have taken place at DTU Compute and Rigshospitalet, with an external stay supervised by Ben Glocker in the BioMedIA group at Imperial College London.

Kongens Lyngby, August 30, 2024

Paula López Diez

Acknowledgements

Firstly and foremost, I would like to thank all my supervisors. Rasmus R. Paulsen, Jan Margeta, and François Patou; thank you for your support and guidance during these years, especially thank you for sticking with me when everything seemed to fall apart and we had to reinvent ourselves. Thank you for selflessly sharing all your knowledge and letting me explore the avenues that I found more interesting. I have learned a lot from you and will always be grateful for your support during this project.

I would also like to thank all the clinical collaborators who have made this project possible, for their patience and passion to share clinical insights with a computer scientist. I really appreciate it. I am particularly grateful to my colleagues at Rigshospitalet, especially the radiology department and the interventional cardiology fellows, for welcoming me with open arms.

I would like to thank Ben Glocker for hosting me on my external stay at Imperial College, and to all the BioMedIA group that made my time in London an unforgettable experience.

Thank you to all my colleagues at Visual Computing, for all the scientific discussions, the lunch breaks, and all the quality time we have shared over the last 3 years. Special thank you to Josefine Vilsbøll Sundgaard and Kristine Aavild Sørensen for showing me the way, being always open to discussing any scientific and personal matter, and for the excellent company in our conference and dissemination trips. I would also like to thank all the students that I have had the pleasure of supervising in diverse projects during these years, for their enthusiasm and contagious thirst for knowledge.

As I prepare to submit my dissertation during this last week of August, with the DTU campus bustling with new students about to embark on their master's journeys as I did six years ago, it feels like the right moment to express my gratitude to the Technical University of Denmark, which provides such a nurturing environment to study and especially to conduct research. I am also immensely grateful to my previous professors and friends from Spain, particularly from Asturias, the region where I grew up, received my early education, and developed my core values. Leaving one's homeland is never easy, but Denmark has been a wonderful home to me these past years, and for that, I am very grateful. The friends I've made in Copenhagen have been crucial to my well-being here, and I want to thank them for their support, the dinners, field trips, runs, Friday beers, and karaoke sessions. You have truly been my family here.

I would also like to express my deep appreciation to my family, who has supported me from afar since I moved to Denmark. They are always present in my heart, and without them, none of this would have been possible.

Last but certainly not least, thank you to Marcos, for being my rock through all these years and for choosing to share his life with me.

List of contributions

All the contributions included in this list have been peer-reviewed.

Main contributions

A **Paula López Diez**, Josefine Vilsbøll Sundgaard, François Patou, Jan Margeta, Rasmus Reinhold Paulsen. Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. *Proc. of Medical Image Computing and Computer Assisted Intervention*, MICCAI 2021. DOI:10.1007/978 − 3 − 030 − 87202 − 1_50 [47]

B Ana-Teodora Radutoiu, François Patou, Jan Margeta, Rasmus Reinhold Paulsen, and **Paula López Diez**. Accurate localization of inner ear regions of interest using deep reinforcement learning, *Proc. 13th International Workshop, MLMI 2022*, Held in Conjunction with MICCAI 2022. DOI:10.1007/978 − 3 − 031 − 21014 − 3_43 [65]

C **Paula Lopez Diez**, Kristine Aavild Juhl, Josefine Vilsbøll Sundgaard, Hassan Diab, Jan Margeta, Francois Patou, and Rasmus Reinhold Paulsen. Deep Reinforcement Learning for Detection of Abnormal Anatomies. *Proc. of the Northern Lights Deep Learning Workshop*, 2022. DOI:10.7557/18.6280 [21]

D **Paula Lopez Diez**, Kristine Aavild Sørensen, Josefine Vilsbøll Sundgaard, Khassan Diab, Jan Margeta, Francois Patou, and Rasmus Reinhold Paulsen. Deep Reinforcement Learning for Detection of Inner Ear Abnormal Anatomy in Computed Tomography. *Proc. of Medical Image Computing and Computer Assisted Intervention*, MICCAI 2022. DOI:0.1007/978-3-031-16437-8_67 [46]

E **Paula López Diez**, Josefine Vilsbøll Sundgaard, Jan Margeta, Khassan Diab, François Patou, and Rasmus Reinhold Paulsen.
Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images. *Journal of Computerized Medical Imaging and Graphics*, 2023. DOI:10.1016/j.compmedimag.2024.102343 [22]

F **Paula López Diez**, Jan Margeta, Khassan Diab, François Patou, Rasmus Reinhold Paulsen. Unsupervised classification of congenital inner ear malformations using DeepDiffusion for latent space representation. *Proc. of Medical Image Computing and Computer Assisted Intervention*, MICCAI 2023. DOI: 10.1109/ISBI52829.2022.9761610 [45]

G **Paula López Diez**, Jan Margeta, Javier Gómez-Herrero, Davorka Lulic, Yannick Willemen, Klaus F. Kofoed, Ole De Backer, and Rasmus R. Paulsen. Peridevice leaks following left atrial appendage occlusion - analysis with morphology descriptive centerlines and explainable graph attention network. *Proc. of 15th International Workshop, STACOM 2024*, Held in Conjunction with MICCAI, 2024. (In proceedings)

Additional contributions

I Jan Margeta, Raabid Hussain, **Paula López Diez**, Anika Morgenstern, Thomas Demarcy, Zihao Wang, Dan Gnansia, Octavio Martinez Manzanera, Clair Vandersteen, Hervé Delingette, Andreas Buechner, Thomas Lenarz, François Patou, and Nicolas Guevara. A Web-Based Automated Image Processing Research Platform for Cochlear Implantation-Related Studies *Journal of clinical medicine*, 2022. DOI: 10.3390/jcm11226640 [52]

II Kristine Aavild Sørensen, **Paula Lopez Diez**, Jan Margeta, Yasmin el Youssef, Michael Pham, Jonas Jalili Pedersen, Tobias Kühl, Ole de Backer, Klaus Kofoed, Oscar Camara and Rasmus R. Paulsen. Spatio-temporal neural distance fields for conditional generative modeling of the heart, 2024. (In proceedings)

III Ana-Teodora Radutoiu, **Paula Lopez Diez**, Jan Margeta, Yasmin el Youssef, Michael Pham, Jonas Jalili Pedersen, Tobias Kühl, Ole de Backer, Klaus Kofoed, Oscar Camara and Rasmus R. Paulsen. Spatio-temporal neural distance fields for conditional generative modeling of the heart, 2024. (In proceedings)

Supervised student projects

- Ana-Teodora Radutoiu. B.Sc. Automatic ROI detection of Inner Ear in CT, 2022
- Theresa Dahl Frehr and Cathrine Underbjerg Hansen. B.Sc. Facial and Chorda Tympani Nerve Landmark Detection and Angle Estimation on MR Images, 2022
- Javier Garcia Ciudad. Special Course M.Sc. Synthetic generation of CT images from other image modalities using deep learning, 2022
- Special Course, Deep learning for medical image segmentation, 2022
- Special Course, Human-in-the-loop and anatomical image segmentation, 2022
- Muhammad Roshan Mughees. M. Sc. Automatic cochlear implant to facial nerve proximity assessment tool using transformer-based segmentation networks, 2022
- Ana-Teodora Radutoiu. Special Course, Human-in-the-loop and accurate landmark localization in medical images, 2023
- Mark Bindesbøll and Rasmus Vilhelmsborg Gøl. B.Sc. Estimating dorsal tilt on distal radius X-Ray images using deep reinforcement learning landmark detection, 2023
- Bjørn Marius Schreblowski Hansen and Mathias Micheelsen Lowes. Special course, Human-in-the-loop and active learning for abdominal organ segmentation, 2023
- Thina L. Thøgersen. M. Sc. Self-Supervised Denoising for Rician Image Data, 2023
- Anna Bøgevang Ekner. B. Sc. Deep reinforcement learning for path tracing in 3D medical images, 2023
- Andreas With Aspe. M. Sc. AI-driven image analysis on computed tomography for 3D vertebral segmentation and detection of osteoporotic complications, 2023
- Special Course, Advanced and user-friendly methods for vestibular analysis in CT data, 2024
- Mark Bindesbøll. Special Course, Segmentation and analysis of the renal arteries in 3D CT scans, 2024

- Caroline Borup Jeppesen. Special Course, Segmentation and analysis of the spine in 3D CT scans, 2024
- Victoria Charlotte Ipsen. M.Sc. AI driven tracing and analysis of the renal arteries in 3D CT scans, 2024
- Bjørn Marius Schreblowski Hansen. M.Sc. AI driven spatiotemporal modelling of the left atrial appendage, 2024
- Mathias Micheelsen Lowes. M.Sc. AI driven spatio-temporal modelling of the heart muscle, 2024
- Julie Kofod Boel and Katja Refsgaard Norsker. B.Sc. AI driven outlier detection of the human vertebra from computed tomography, 2024

List of Abbreviations

3D three-dimensional

ABI auditory brainstem implantation

AF atrial fibrillation

AI artificial intelligence

AuC area under the curve

CAE convolutional autoencoder

CI cochlear implant

CN cochlear nerve

CNN convolutional neural network

CT computed tomography

DL deep learning

DQN deep Q-network

DRL deep reinforcement learning

ENT ear-nose-throat doctor

FN facial nerve

FNS facial nerve stimulation

FO fossa ovalis

GAN generative adversarial network

GAT graph attention network

HA hearing aid

IAC internal acoustic canal

IEM inner ear malformation

IVC inferior vena cava

LA left atrium

LAA left atrial appendage

LAAO left atrial appendage occlusion

LV left ventricle

MDP Markov decision process

MRI magnetic resonance imaging

PDL peridevice leak

PDM point distribution model

RA right atrium

RL reinforcement learning

ROI region-of-interest

SNHL sensorineural hearing loss

TS TotalSegmentator

VAE variational auto-encoders

Contents

CHAPTER 1

Introduction

1.1 Motivation

Medical imaging has significantly advanced the diagnosis and treatment of various medical conditions, enabling detailed visualization of internal anatomical structures. Developing advanced computer vision methods to analyze and extract the most information out of these images is a very active field of research. There are many different research avenues that can facilitate or improve the daily workflow of clinicians working with these types of images. In our work, we focus on the development of deep learning (DL) tools to assist clinicians in their preoperative clinical analysis of patients who will receive a medical implantable device. More specifically, the main part of our work focuses on cochlear implant (CI) therapy which is used to restore the hearing capacities of patients who suffer from profound hearing loss. The second part of our work involves left atrial appendage occlusion (LAAO) which is a procedure used as an alternative prevention treatment for patients with a high risk of developing blood clots and consequently suffering a stroke. Both of these implantable devices present unique challenges due to the intricate and variable anatomies where they must be inserted. Computer-based methods applied to medical images for the therapy of these implantable devices should consider the various clinical challenges of these procedures as well as the functionality of the therapy itself.

In the context of CIs, accurate diagnosis and treatment planning are heavily dependent on a detailed understanding of the inner ear's anatomy. However, the complexity of this region has historically limited the application of advanced image analysis techniques. Similarly, cardiac procedures like LAAO involve navigating complex anatomical structures that are difficult to study and characterize mostly due to the great inter-patient morphological variability and the novelty of the therapy itself.

The motivation behind this thesis is to address these gaps by developing novel methods in three-dimensional (3D) image analysis, tailored specifically for the unique challenges posed by these critical anatomical regions. By leveraging cutting-edge techniques such as deep reinforcement learning, unsupervised methods, and graph attention networks, the research aims to enhance the use of medical imaging analysis, ultimately supporting clinicians in making more informed decisions and improving patient outcomes.

1.2 Thesis objectives

The primary objective of this thesis is to develop and validate innovative 3D image analysis methods that address the specific challenges associated with clinical interventions involving implantable devices. The objectives are pursued with the goal of advancing the technical capabilities of medical imaging while also addressing key clinical challenges. The research focuses on three main areas:

- Landmark detection and characterization: Develop and apply deep reinforcement learning techniques to extract and utilize anatomical landmarks in the inner ear. This will enable more accurate characterization of neural structures and head orientation in computed tomography (CT) scans, which are critical for cochlear implant procedures.
- Anomaly detection: Propose and evaluate new methods for detecting congenital malformations in the inner ear, these cases present an added challenge for clinicians, especially in the context of cochlear implant therapy. These methods are based on novel and advanced machine learning

techniques and particularly focus on detecting anomalies with a parametric approach that targets local structures more than overall image appearance.

- Shape-specific image analysis: Create shape- and graph-based models to analyze and derive insights from 3D shape data. More specifically classification of congenital malformations based on their cochlear topology and the complexity of LAAO cardiac procedure. These models will capture the relevant anatomical features and their influence providing insights that can inform clinical practice.

1.3 Thesis structure

This thesis is structured in six chapters and seven appendices.

The first chapter presents a high-level introduction of the thesis motivation and objectives as well as a description of how the final manuscript is structured. The second chapter introduces the clinical and technical background, as well as the state of the art, that contextualizes the different contributions. In the third chapter, all the data used in this work, which consists of different 3D CT scan images, is introduced. Chapter four presents all the different contributions and how they relate to each other, analyzing the different outcomes and the research flow during the PhD program. This chapter should be considered a brief overview of the contributions, which are also discussed in chapter five, for a full description of the methods and results the reader is referred to the publications which are available in Appendix A-G.

CHAPTER 2

Background

This chapter presents the clinical and technical background essential for understanding the contributions of this thesis. The first section explores the clinical aspects of two implantable devices: the cochlear implant and the left atrial appendage occluder. We examine the functionality of these devices, the anatomical considerations in their surgical placement, and the utilization of CT medical imaging for patients undergoing these procedures. The second section provides the technical background and state of the art, offering the necessary context for the DL methods developed in this thesis.

2.1 Clinical background: implantable medical devices

Implantable medical devices have revolutionized modern medicine, offering life-saving interventions and significantly improving the quality of life for many patients. Devices such as pacemakers, defibrillators, neurostimulators, cochlear implants, and insulin pumps integrate advanced technology with clinical practice to support individuals with chronic conditions.

The process of implanting a medical device involves several critical stages to ensure patient safety and optimal outcomes. These stages include medical evaluation, diagnosis, therapy decision, surgical planning, surgical procedure, and postoperative follow-up. The methods developed in this thesis are specifically designed to enhance preoperative tasks. Our focus is on creating automated techniques that assist clinicians during the diagnostic stage and provide valuable insights for surgical planning, thereby adding value and reducing the overall workload.

In the following sections, we describe the clinical aspects of the two types of implantable devices involved in this thesis, to understand the contributions from a medical standpoint. We focus on an active implantable device, the CI, and a passive one, the left atrial appendage occluder.

2.1.1 Cochlear implants

Hearing is a crucial sense for humans. As social beings, our interactions usually involve conversation, which is essential for our development. Additionally, hearing is important for interacting with our environment, as it warns us of danger and plays a vital role in human survival. Partial damage or the complete loss of this sense can significantly limit our experiences and cognitive development.

CIs have been developed to restore the hearing ability of patients with congenital or acquired severe to profound sensorineural hearing loss (SNHL). The implant comprises an external portion and an internal portion placed under the skin as shown in figure 2.1. It consists of:

- A microphone to capture sound.
- A sound processor to select and process the captured sounds.
- A transmitter that sends the processed sound to a receiver, placed under the skin, which converts the signal to electrical impulses.
- An electrode array to deliver the electrical impulses to various regions of the cochlea.

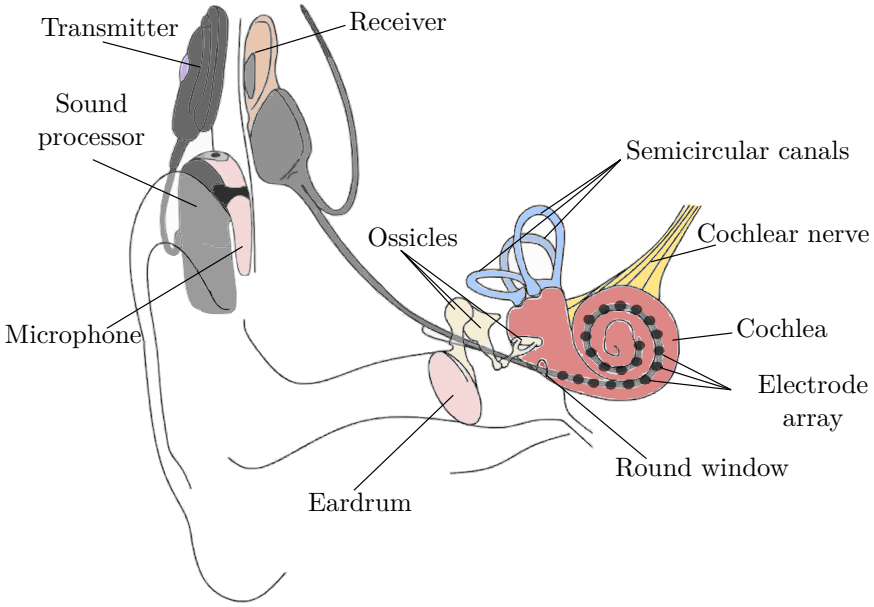


Figure 2.1: Cochlear implant workflow. The microphone captures sounds and sends the signal to the sound processor which transforms the sound signal and passes it to the transmitter which sends the signal to the receiver that's been put under the skin behind the ear. The receiver converts the signal to electrical impulses. These signals activate the cochlear nerve in the right frequencies so the brain can hear the signals as sounds.

2.1.1.1 CT images for CIs

An otorhinolaryngologist, also known as ear-nose-throat doctor (ENT), is responsible for determining the definitive indication for CI therapy based on a comprehensive interdisciplinary diagnostic work-up. This work-up encompasses audiological, radiological, psychological, and clinical assessments [17].

This dissertation focuses on the radiological aspects, particularly on CT imaging of the inner ear. The inner ear is a closed and fluid-filled tunnel system encased within the temporal bone which is also referred to as the bony labyrinth due to the complexity of the anatomy in this region. It comprises three primary components: the cochlea, the vestibule, and the semicircular canals.

High-resolution CT imaging of the petrous part of the temporal bone is utilized to confirm the presence of the cochlear nerve (CN), assess the fluid content within the cochlea, and exclude the presence of chronic ear diseases or malformations [17]. This research specifically aims to characterize the neural structures in this region, with an emphasis on the cochlear and facial nerves, and to detect and characterize congenital malformations of the inner ear.

2.1.1.2 Facial and cochlear nerves

CI surgery typically involves percutaneous cochlear access, meaning a single hole is drilled from the skull surface to the cochlea to insert an electrode array uniformly into the cochlear structure. Planning the trajectory, that passes through the narrow facial recess, is one of the critical points of this procedure. The damage of the facial nerve (FN) has direct consequences such as dry eye, reduction of saliva flow, and loss of taste from the anterior two-thirds of the tongue and the palate [81].

Post-implantation FN electrical stimulation is also a well-known complication of CI therapy that has an estimated reported rate was approximately 6% [83] and a significantly higher rate (reaching almost 50%) in patients with otosclerosis, cochlear malformation, or ossification [82]. Facial nerve stimulation (FNS) is caused by the electric current, passing through the electrode of the implant to

the spiral ganglion cell, spreading to the nearby FN causing symptoms ranging from simple awareness to severe facial spasms [8]. FNS can frequently be resolved with small changes in speech processor fitting but, sometimes, this can lead to a reduction in the outcome. In some cases, FNS results in such severe discomfort or limited stimulation range that the CI becomes useless, which leads to explantation and reimplantation. Several factors contribute to the etiology of FNS following cochlear implantation, where it has been shown that the configuration of the stimulation parameters is one of them [27]. The proximity of the labyrinthine segment of the FN to the superior segment of the basal turn of the cochlea has also been shown to be a notable cause of FNS [37, 9]. This motivates the need to characterize the FN in the region adjacent to the cochlea, where the complex and intricate morphology, shown in Figure 2.2, presents significant challenges.

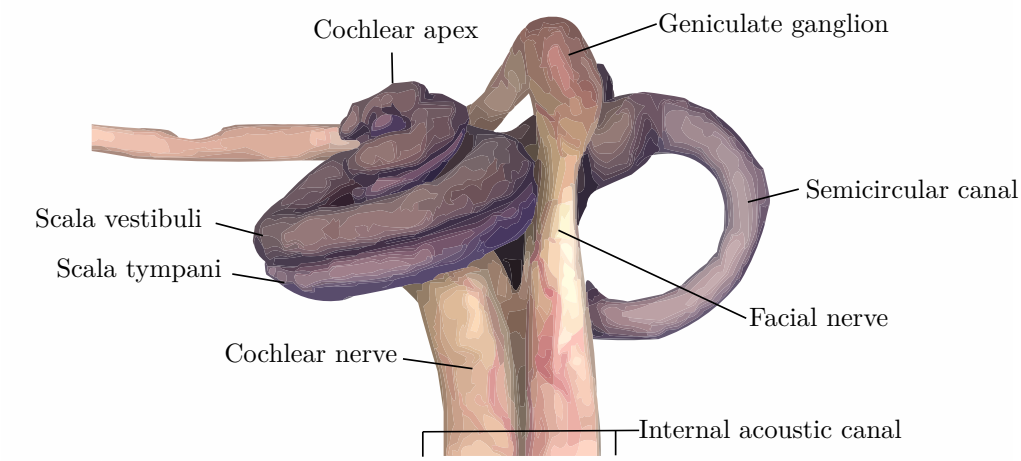


Figure 2.2: Anatomical 3D visualization of the cochlea and semicircular canals together with both the facial and the cochlear nerve. Note the proximity of the cochlear structure, especially to the scala tympani which is the cavity where the electrode array of the implant is placed.

The CN is responsible for transmitting auditory information to the brain. There is a large range of pathologies affecting the cochlear nerve, which are usually analyzed with magnetic resonance imaging (MRI). This is often the imaging of choice for the investigation of pathologies of the cochlear nerve because MRI has a better contrast for soft structures, such as nerves. However, the use of this imaging modality presents two main challenges: the resolution is usually not sufficient for such small structures as the CN [18] and the MRI needs a longer time to process. The latter challenge requires the patient to lie still for a long period, which, in the case of small children, often leads to sedation with all the risks associated. Children are the main recipients of CI therapy, on average, 78% of deaf children receive cochlear implants, whereas only 6.6% of adult CI candidates receive them [19]. Utilizing pre-operative CT scans, which can be acquired quickly, to detect cochlear nerve deficiency (CND) — one of the various causes of hearing loss — would be both beneficial and time-saving.

CND encompasses conditions such as CN aplasia (complete absence of the CN), absence of the CN in the internal auditory canal, and CN hypoplasia, characterized by a reduced CN size within the internal auditory canal. Typically, the diameter of the CN is reported to be similar to or slightly larger than that of the FN. When the CN is significantly smaller than the FN in the internal acoustic canal (IAC), it is classified as CN hypoplasia [33]. Deficient CNs have been associated with poor cochlear implant performance. Therefore, the integrity of the CN is usually assessed before selecting CI therapy, as its proper functionality is crucial for achieving a successful outcome. Visualization of the cochlear nerve in the fluid-filled internal auditory canal is only possible in MRI images. However, the width of the canal itself, which has been linked with CN deficiency, can be analyzed in CT images [96].

2.1.1.3 Congenital inner ear malformations (IEMs)

IEM with associated SNHL is a major cause of childhood disability. IEM has been reported with an incidence of 20%–30% among children with congenital hearing loss [12] where the diagnosis is usually based on a radiological examination of a CT image of the inner ear. For the remaining cases of congenital hearing loss, the pathology involves deficient inner ear hair cells and CT images reveal normal findings. Detecting and identifying such IEMs from standard imaging modalities is a complex task, even for expert clinicians, given the complexity of the anatomy and the great anatomical variation among malformations. Studies have defined several categories for these malformations, such as the one proposed by Sennaroğlu [76] which is one of the most popular works used for classifying congenital malformations of the inner ear. An overview of the types of malformations as well as the radiological findings , including FN anomaly, is shown in Table 2.1.

Each type of IEM is unique and the treatment modality available is different depending on the type of malformation and sometimes even in the individual features present in each specific patient. Some cases may be managed by a hearing aid (HA), others need CI, and some cases are candidates for auditory brainstem implantation (ABI), an overview of the possible treatment available for each type of malformation is shown in Table 2.1. Cases presenting congenital IEMs raise many challenges during the planning and execution of CI surgery, often necessitating the surgeon to discover and adapt as the procedure progresses.

IEMs are increasingly understood, both in terms of their morphological characteristics and the development of rehabilitative options. What was once considered intractable malformations with little hope for rehabilitation are now conditions that can be effectively managed in specialized centers [12]. Notably, research has shown the correlation of the morphological characteristics of these malformations with specific stages of developmental arrest in the embryo [68] suggesting that the deformities result from an arrest of developing during varying stages of inner ear organogenesis which are shown in Figure 2.3. It is therefore known that the malformations are present in a continuous spectrum and that their clinical classification in discrete clinical terms is difficult. A sketch of the conceptual understanding of different types of malformations is shown in Figure 2.4.

IEM	Radiology findings	FN anomaly	Possible Treatment
Complete labyrinthine aplasia	Absent Labyrinth	Yes	ABI
Rudimentary otocyst	Incomplete otic capsule remnant	Yes	ABI
Cochlear aplasia (CA)	Absent cochlea	Yes	ABI
Common cavity (CC)	Round or ovoid cystic structure for cochlea and vestibule	Yes	CI or ABI
Cochlear hypoplasia (CH)	Cochlear size small (type I, II, III and IV)	Yes	HA, CI or ABI
Incomplete partition I (IP-I)	Cystic cochlea	Possible	CI or HA
Incomplete partition II (IP-II)	Cystic cochlear apex	Not expected	CI or HA
Incomplete partition III (IP-III)	Modiolus absent, interscala septa present	Yes	CI or HA
Enlarged vestibular aqueduct	Normal cochlea with enlarged vestibular aqueduct	Not expected	CI or HA
Cochlear aperture abnormalities	Narrow or absent cochlear aperture	Not expected	Depending on the CN, CI or ABI

Table 2.1: Classification of IEMs according to [76] with their corresponding basic radiological findings, FN anomaly presence, and available treatment options.

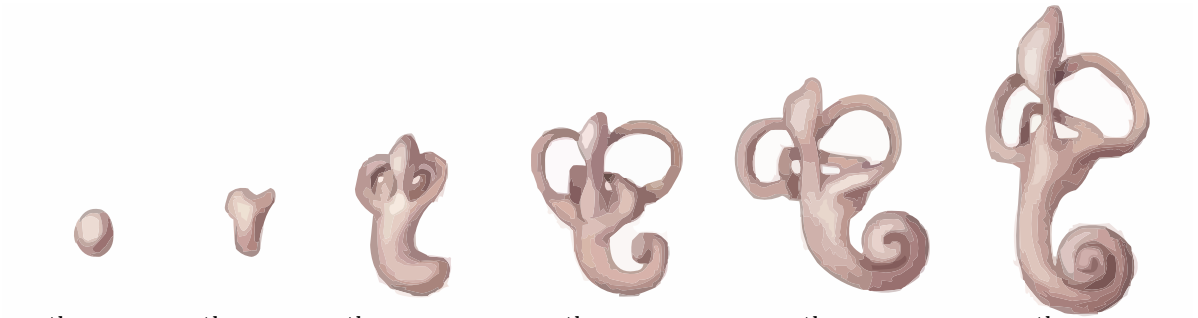


Figure 2.3: Embryogenesis of cochlear malformation based on [68].

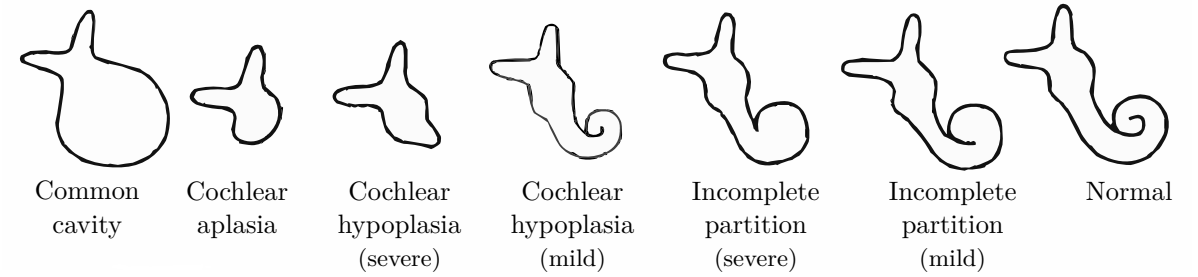


Figure 2.4: Schematic representation of cochlear malformations based on [36].

The pathogenesis of IEMs has been the subject of various hypotheses, with genetic factors frequently implicated. Recent advancements in genetic consultation and testing have yielded new insights into the potential connections between IEMs and genetic anomalies. In parallel, the evolution of neuroradiological imaging techniques, including advanced CT and MRI has provided valuable data, particularly in syndromic patients. In certain syndromes, hearing loss is not only well-documented but also linked to specific IEMs. The identification of a genetic mutation is crucial for assessing the risk of recurrence in future pregnancies and may prompt further neuroradiological studies of the hearing system. Conversely, the detection of IEMs can serve as an indicator for genetic testing and the exploration of extra-auditory features that may suggest an underlying syndrome [68].

Besides the research done for finding optimal categories for the different IEM pathologies and getting a deeper understanding of them from a clinical standpoint, different clinical guidelines for their detection and classification based on CT images have been proposed. Due to the great complexity of giving a diagnosis from the CT image, different approaches based on the visual exploration of CT scans involving explicit measurements and humans' natural ability for pattern recognition have been proposed like the method presented in [20].

2.1.2 Left atrial appendage occluder

The heart is a vital organ responsible for pumping blood throughout the body, which involves supplying oxygen and nutrients while removing carbon dioxide and other waste products. An anatomically normal heart consists of four chambers: two atria and two ventricles, an illustration of the heart anatomy is shown in Figure 2.5. The left atrium (LA) plays a crucial role in receiving oxygenated blood from the lungs and pumping it into the left ventricle (LV), which then distributes it to the rest of the body.

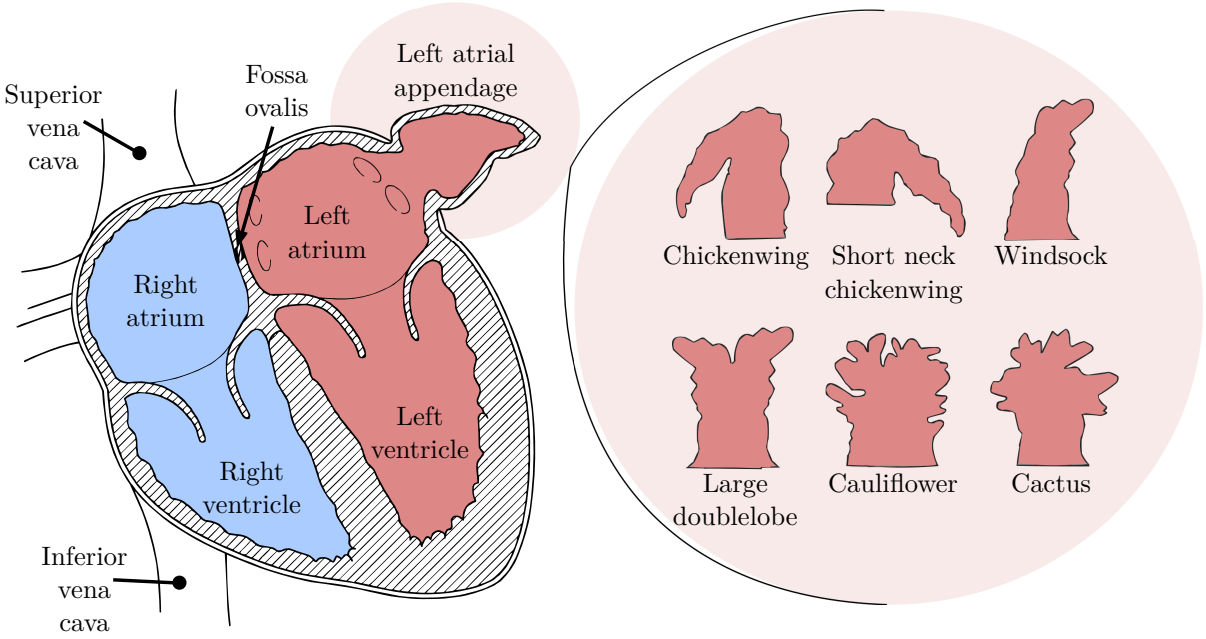


Figure 2.5: Illustration of the heart anatomy on the left including the location of the left atrial appendage (LAA) and on the right an illustration of the different morphological categories of this anatomical structure.

Atrial fibrillation (AF) is a common cardiac arrhythmia characterized by an irregular and often rapid heart rate. This condition increases the risk of thrombus formation[49], particularly in the left atrial appendage (LAA) in which around 90% of intracardiac thrombi are found for these patients [98]. The LAA is a small, ear-shaped pouch in the LA that has been found to have significant variations in shape and size among humans as shown in Figure 2.5. A thrombus created in this region can travel to the brain and cause an ischemic stroke, which is a severe medical emergency due to the interruption of blood flow to the brain. After an ischemic stroke, mortality is approximately 10% in the first thirty days and 40% at the end of the first year [55]. Survival is dependent on receiving early treatment.

In recent years, the LAAO has emerged as an alternative to anticoagulation therapy for stroke prevention in patients with AF. The LAAO device mechanically seals off the LAA, thereby preventing thrombus formation within this high-risk area. This approach offers a potential solution for patients who are unsuitable for long-term anticoagulation or those who experience significant bleeding complications.

2.1.2.1 Imaging and surgical procedure

LAAO is a complex and multi-step transcatheter procedure that requires precise coordination and advanced imaging techniques. The procedure begins with the insertion of a femoral venous sheath, through which a catheter is advanced to the right atrium (RA). A transseptal puncture through the fossa ovalis (FO) is then performed to gain access to the LA, a delicate step requiring exact positioning to avoid complications as can be seen in Figure 2.6. Real-time imaging, typically using transesophageal echocardiography and fluoroscopy, guides the catheter to the LAA. The occluder device, pre-loaded in the catheter, is then carefully deployed by retracting the sheath. Accurate placement of the device is critical; it must seal the LAA completely without interfering with adjacent structures.

Given the complexity and expertise needed for this procedure, which, combined with the high heterogeneity present in the morphology of this specific anatomy, makes a detailed preoperative analysis always needed. This analysis often involves a cardiac CT image which provides detailed anatomical information about the LAA, allowing for precise measurements of its size and shape, which are essential for selecting the appropriate device type and size.

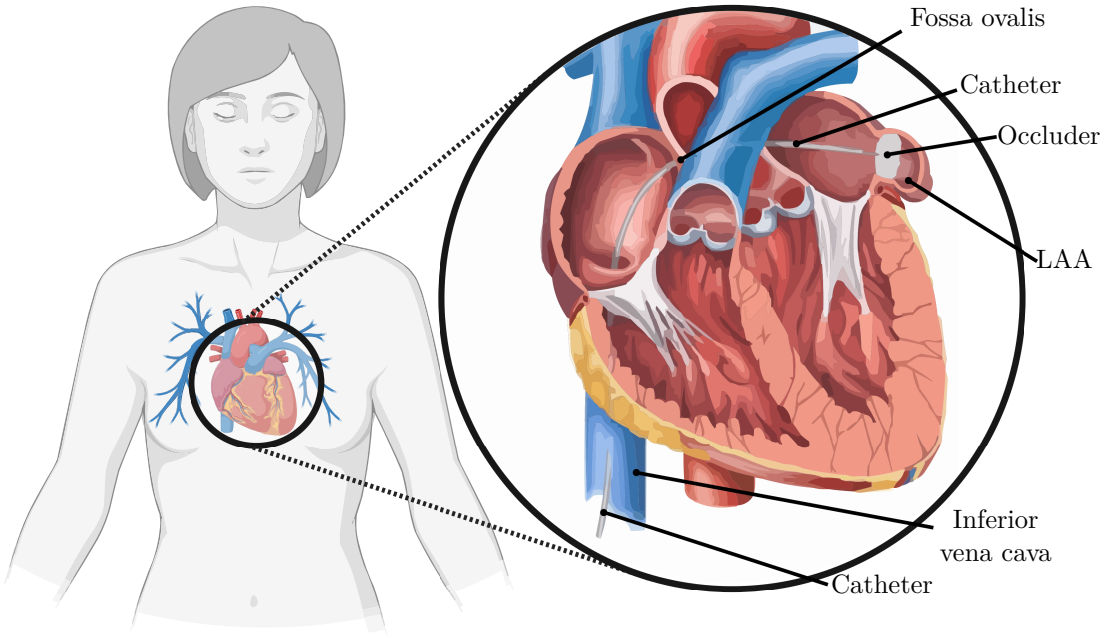


Figure 2.6: LAAO procedure illustration showing the catheter trajectory, the occluder, and the anatomical points of interest. The catheter goes from the inferior vena cava (IVC), into the RA, then through the FO into the LA where it has direct access to close the opening between the LA and the LAA with the device.

Different types of devices and sizes are available in the market. We can differentiate two main categories based on the occluder mechanism with a single-occlusive plug type mechanism, and devices with a dual-occlusive technology consisting of a lobe to fill and anchor in the cavity of the LAA and a disc to seal the LAA orifice. During the procedure, real-time imaging with contrast is used to determine whether or not the placement of the device seems optimal before the device is released from the sheet.

After the procedure, the patient is monitored for a few hours and is often soon released from the hospital. The correct device positioning is checked by extracting cardiac CT images with contrast after the process of endothelialization has been completed, usually around 3 to 5 months after the procedure.

2.1.2.2 Peridevice leak (PDL) following LAAO

PDL is a significant concern following LAAO procedures, occurring when there is residual flow around the occluder device into the LAA, which was documented in 32% of patients at 1-year follow-up [87]. This can undermine the effectiveness of the procedure by allowing blood to enter and stagnate in the LAA, thus perpetuating the risk of thrombus formation and subsequent embolic events.

PDLs have yet not been studied in depth given LAAO is a relatively new therapy and their multi-factorial nature. Several factors can contribute to the presence of PDLs following LAAO including procedural factors, LAA anatomy, implanter expertise, device design, and final device positioning [40]. The great morphological heterogeneity of the LAA anatomy, combined with the fact that most patients who undergo this procedure present other severe pathologies, makes finding features that might link with the prognosis of the procedure in terms of PDL presence more challenging.

Correctly characterizing a PDL requires an exhaustive analysis of the 3D CT image with contrast in the LA. The presence of contrast patency in the distal LAA anatomy is indicative of a leak. To determine whether the leak is a PDL or not, a contrast gap adjacent to the device disc and/or in the lobe must be observed. If there is no visible gap, it can be due to a contrast leakage through the device fabric, as seen with incomplete endothelialization, or a microleak, both of these cases are not

categorized as PDLs. An illustration showing the different types of leaks and their corresponding PDL classification is displayed in Figure 2.7.

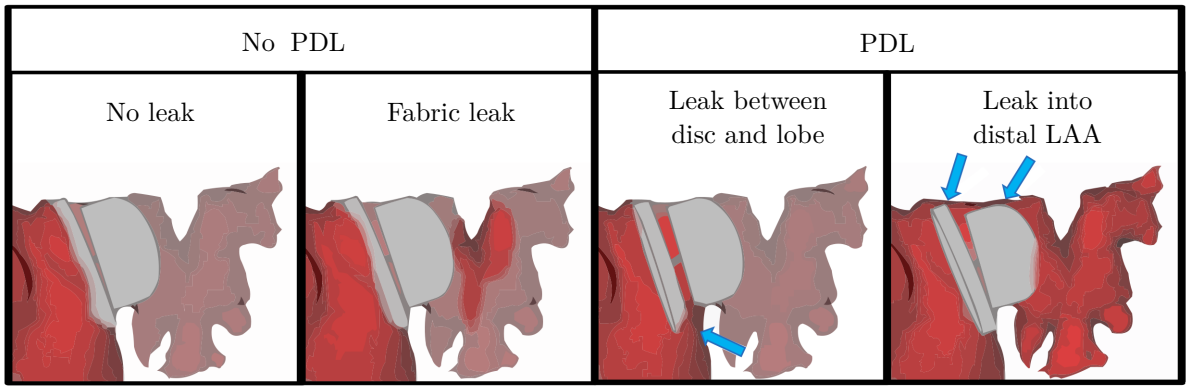


Figure 2.7: Illustration of types of device leak following LAAO procedure and their classification into PDL presence or not.

2.2 Technical background and previous work

In this thesis, a great variety of DL methods have been used to tackle complex technical problems characterized by limited datasets and clinically challenging questions. Our work revolves around DL techniques for 3D medical image processing such as landmark detection, anomaly detection, unsupervised classification, and explainable graph neural networks. Providing the technical background of such a broad spectrum could be extensive. Therefore we assume the reader has general knowledge of DL and we present the synthesized technical background to provide the high-level contextualization needed to understand the contributions of this dissertation. In this chapter, we give brief technical introductions to the methods used and the contextual information needed to better understand the contributions from a technical perspective which involves deep reinforcement learning (DRL) (utilized in Papers *A*, *B*, *C*, *D*, and *E*), anomaly detection (regarding Papers *C*, *D* and *E*) and 3D shape classification (referring to Papers *F* and *G*). In this section, we also include information about the state of the art of each field that we have incorporated together with the contextual description.

2.2.1 Deep reinforcement learning

The objective of the machine learning technique known as reinforcement learning (RL), which was influenced by cognitive science, is to determine the optimal strategy to solve a given problem. This is accomplished by allowing an agent to learn how to behave in an environment where the only form of feedback is a reward signal.

The agent interacts with an environment E through a cycle of observations, actions, and rewards. At each time step, the agent observes the current state s and selects an action $a \in A$, where A is a finite set of allowable actions. The environment then provides a signal or reward R, which the agent receives and interprets as the outcome of the action a. This feedback loop helps the agent learn and adapt its behavior over time. A schematic representation of this process is shown in Figure 2.8, illustrating the dynamic interplay between the agent and the environment.

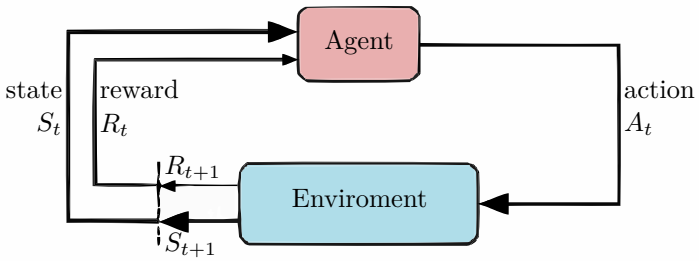


Figure 2.8: RL cycle. Showing the interaction between the agent and the enviroment in a time-step t.

RL occupies a unique position between supervised and unsupervised learning. The RL paradigm involves addressing sequential decision-making problems, where the learning process is guided by limited feedback. All in all, the reward signal guides the agent towards taking correct actions and penalizes it otherwise. Thus, the agent ends up learning a policy π from high-dimensional input. Usually, in complex real-life scenarios, agents would not have total knowledge of all the possible states of the environment because it is not fully observable. This is commonly referred to as a partially observable Markov decision process (MDP).

Numerous problems have been effectively modeled using a MDP. In fact, MDPs have become the standard framework for learning in sequential decision-making scenarios. In an MDP, the environment is represented as a set of states and actions that can be used to manipulate the system's state. The objective is to control the system to maximize a specific performance criterion. It can be defined as a tuple: S, A, O, P, R, Z, and γ, where:

- S - finite set of states.

- A - finite set of actions.
- O - finite set of observations.
- P - state transition probability matrix defined as:

$$P^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a] \tag{2.1}$$

- R - reward function, defined as:

$$R^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a] \tag{2.2}$$

- Z observation function, defined as:

$$Z^a_{s'o} = \mathbb{P}[O_{t+1} = o | T_{t+1} = s', A_t = a] \tag{2.3}$$

- $\gamma \in [0, 1]$ - discount rate

Typically, the reward is influenced not just by the most recent action but also by earlier decisions. To address this, both immediate and future rewards must be considered, which is done using the discount rate γ. The discount rate diminishes the value of future rewards. If the value of γ is set to 1, there will be no discount, which is suitable for deterministic environments where the same actions lead to the same rewards. On the contrary, if $\gamma = 0$, only immediate rewards will be taken into consideration, resulting in a very short-sighted strategy.

One of the value functions used to estimate the expected return of a given state is the quality function $Q^\pi(s, a)$. This function outcomes the expected return from estate s when selecting action a following policy π. It can be defined as [3]:

$$Q^\pi(s, a) = \mathbb{E}_\pi[R | s, a] \tag{2.4}$$

Therefore the best possible policy given Q^π, π', can be derived by greedily choosing the action, a, that maximizes $Q^\pi(s, a)$:

$$\pi'(s | Q^\pi) = \underset{a}{\operatorname{argmax}} \mathbb{E}[Q(s, a)] \tag{2.5}$$

and the optimal policy, π^*, as:

$$Q^*(s, a) = \max_\pi Q^\pi(s, a) \quad \forall s \in S \quad \forall a \in A(s) \tag{2.6}$$

Many methods have been developed to compute good or optimal policies for problems modeled as an MDP among which Q-learning is a popular choice. In Q-learning a policy is learned by estimating the Q-function defined in 2.4. To do so, the Markov property is exploited and the Q-function is transformed into a Bellman equation, with the following shape:

$$Q(s, a) = \mathbb{E}_{s'}[r + \gamma Q(s', \pi(s'))] \tag{2.7}$$

Therefore the value of Q can be improved based on the current estimate:

$$Q(s, a) \leftarrow Q(s, a) + \alpha\delta \tag{2.8}$$

Where α is the learning rate and $\delta = Y - Q(s, a)$ is the temporal difference error, where $Y = r + \gamma \max_{a'} Q(s', a')$ is the target value, allowing to formulate the problem as a dynamic programming problem.

In this thesis, we have used DRL which implies using DL architectures to solve the previously stated problem. In order to find the optimal policy of the MDP, DRL can be used to approximate the optimal function by iteratively sampling a set of states, actions, and rewards. Recent advances in training deep neural networks are used to develop an artificial agent, named a deep Q-network (DQN) [54]. This

network can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning.

DRL is based on training artificial neural networks to approximate either the value function (Q-function), the policy(π), or a combination of both. In other words, DL is used to transform the observation perceived by the agent in the current state s to an action a. During the training process of the DQN the reward r generated by the action a, in that certain state s, is directly related to the loss function that will make our model converge, so the agent can learn the policy π and apply it when the reward function of a certain action is unknown.

Due to the success of Q-learning as a surrogate for the optimal function, we used the Q-learning strategy for the task of landmark detection with the multi-agent DRL framework MARL, proposed in [88], and its communicative version, C-MARL, proposed in [43]. For landmark localization in 3D images, the problem is defined as the environment, E, being the 3D image and the agent is a physical location within the image. The state, s, is a patch of the image centered in the agent's location, and the action set, A, is the movement in one of the six Cartesian directions (up, down, left, right, forward, and backward). The reward, r, is defined as the difference between the distance to the target landmark's location after the last action and in the previous state, meaning it is a positive value if the agent is moving closer and a negative one if it has moved away from the target landmark's location. A sketch of the C-MARL architecture is shown in Figure 2.9 which consists of a common convolutional neural network (CNN) that extracts the relevant features of the current state and a set of fully connected layers for each agent that outputs the estimated Q-value of each of the possible actions A. We can observe that the agents have implicit communication as they all share the same CNN layers and they also have explicit communication by sharing average values of their individual fully connected layers.

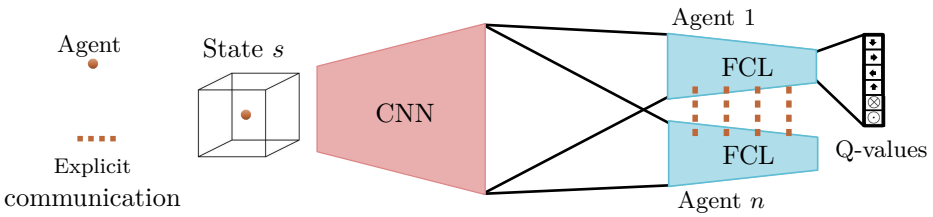


Figure 2.9: Diagram of the C-MARL architecture. The input is a patch centered in the current agent's position (state). We show the CNN which extracts the relevant features of a certain patch. Those features are then passed to each corresponding agent, which consists of a set of fully connected layers that map those features to the estimated expected reward (Q-values) of each of the possible actions (up, right, left, down, forward, or backward).

2.2.1.1 Landmark localization

Landmarks are crucial for medical image analysis techniques. On many occasions, authors rely on them as seeds for other methods (as in [89] and [26]), becoming one of the first and more determining stages of the image processing pipeline. Due to this relevance, many approaches rely on manual extraction, which not only slows down the process but also introduces subjectivity.

Efforts to automate this process have employed conventional machine learning techniques, ranging from classifying whether a voxel contains a certain landmark to determining exact landmark positions through a combination of classification and regression. However, these algorithms are mainly built on handcrafted features and careful design, which could be time-consuming and error-prone [57].

Recently, CNNs have shown promising performance for classification and segmentation tasks in medical imaging due to their capability of learning discriminant feature descriptors from raw images [32]. There are different DL approaches for finding landmarks in medical images, they can be divided

into two categories [51]. One where the 3D image is treated as an input and the coordinates of the landmarks are the output, basically consisting of an end-to-end approach as in [30, 59, 109]; and a second one in which the dimension on the input is reduced using a 2 or 2.5D as a representation of the volume and the landmarks are detected separately as in [108, 42, 44]. The first approach requires a bigger computational effort as the input has 3 dimensions but it also keeps the spatial relationship between landmarks into account and avoids extra preprocessing steps.

Within the DL field, another differentiation between approaches can be done: the ones using conventional DL or the ones based on DRL.

For the conventional DL, there have been many applications with different structures and models that have shown great performance. For example, the one implemented by Zheng *et al.* [109] in which the extracted patches from 3D images were combined with two cascaded multilayer perceptrons using the volumetric classification to locate the landmarks. Also, Zhang *et al.* [108] published an approach where one CNN was first used for the inference of the 3D displacement vectors between landmarks and input patches and then followed by another CNN to model the correlations among image patches. We can also mention the work of Payer *et al.* [59] which used one CNN (U-Net [69]) to predict all landmarks using heatmaps, these were generated by evaluating a certain landmark and its relative position to the others. Later Zhang *et al.* [108] proposed a two-stage task-oriented method in a cascaded manner, which allows real-time detection of large-scale anatomical landmarks on the brain and prostate images.

When talking about the RL approaches, Ghesu *et al.* in 2016 [29] used an RL agent to navigate through a 3D image with fixed step actions chosen by learning to follow the optimal path from any location to the landmark by maximizing the accumulated rewards of the different steps. In 2017, Xu *et al.* [99] developed a supervised method that classifies actions using image partitioning techniques, the model extracts an action map for each voxel of the input image across the whole volume (directional classes towards the target point) using a CNN, due to the computational complexity of 3D CNNs this approach is restricted to small-sized 3D images [2]. This cost has been reduced thanks to the patch-based iterative CNN approach proposed in 2018 [44, 57] which can be adapted for unique or multiple landmark detection. To exploit multi-scale image representation, Ghesu *et al.* extended their RL-based landmark detection in [30, 28].

There have been different approaches for landmark localization in the context of cochlear implant research, particularly focusing on the inner ear. Zhang *et al.* [107] proposed a random forest-based method for automatically localizing anatomical landmarks within this region. Their approach employs regression to identify potential landmark candidates, which are then refined using a heuristic technique based on prior knowledge of spatial relationships between landmarks. Heutink *et al.* [34] utilized three separate CNNs to locate the cochlea in ultra-high-resolution CT images. They applied a sliding window technique with each trained CNN and averaged the resulting probability maps to determine the cochlea's position. Wang *et al.* [94] achieved precise automatic localization of cochlear landmarks using just one annotated image. Their method involves localizing a structure of interest, estimating its location in the target image using a CNN, and applying a non-rigid registration algorithm to transfer the landmarks to the target image. Additionally, Wang *et al.* [93] introduced a technique for locating three anatomical landmarks in the cochlea by combining a simple localization algorithm with a self-designed 2D CNN-based multi-class classification network and a sliding window approach. In the work proposed by Margeta *et al.* [53], they use a 3D U-Net [70] architecture to locate three landmarks in both pre- and post-operative CT scans of CI patients and use them to estimate the canonical pose of the cochlea.

2.2.2 Anomaly detection

One of the most interesting and necessary clinical applications for processing medical images is anomaly detection. Building reliable systems that can detect and raise a flag when the processed image does not belong to the healthy subjects group is crucial to improving the screening process. This could support radiologists in their daily workflow and potentially speed the processing times and consequently the waiting time of patients as well as prevent overseeing potential pathologies. Despite its importance in

clinical applications, the detection of anomalies in medical images is still a very challenging task and a very active research field, mostly due to the rarity, sparsity, heterogeneity, and great diversity of anomalies.

Recent advancements in machine learning have enabled the development of automated medical image analysis methods that exhibit exceptional performance in detecting anomalies. In the context of anomaly detection, there are three primary setups: supervised, semi-supervised, and unsupervised anomaly detection. Supervised anomaly detection is analogous to classification using a highly imbalanced dataset. Despite their impressive performance, these methods, which primarily rely on supervised DL, present several limitations. Firstly, they require large and diverse annotated datasets for training, which are scarce and costly to obtain. Secondly, the models generated are restricted to identifying lesions that are similar to those found in the training data, a significant drawback for rare diseases where collecting sufficient training data is particularly challenging. Furthermore, the bias of these methods towards expected anomaly distributions constrains them to not be applicable beyond the specific pathology they have been trained on, which narrows the scope of detectable pathologies potentially overlooking other pathologies in medical imaging.

Semi-supervised anomaly detection focuses on training a model using only one class, typically the normal (healthy) class, and then applies the model to both healthy and pathological data, reporting the corresponding anomaly scores. Unsupervised anomaly detection, on the other hand, utilizes both normal and anomalous data without employing labels. Instead, it operates purely on the intrinsic properties of the dataset, such as distances or densities, to detect anomalies. However, semi-supervised methods are frequently mischaracterized as unsupervised in the literature, leading to the interchangeability of these terms and potential confusion.

In medical imaging, unsupervised anomaly detection revolves around the discovery of the unexpected, aiming to identify deviations from standard patterns through statistical methods. This technique hinges on detecting outliers—data points that significantly differ from the rest of the dataset and are thus flagged as anomalies. For these algorithms to be effective, they must be designed to generalize across various datasets and remain independent of specific pathologies, allowing them to detect a broad range of previously unseen anomalies. To achieve this, algorithms are trained on extensive datasets that capture a wide spectrum of normal variations, enabling the identification of outliers without the need for labeled data specific to any particular condition. The crux behind most of the semi-supervised and unsupervised methods for anomaly detection in medical images is faithfully modeling the healthy anatomy with unsupervised deep (generative) representation learning. Therefore, the methods leverage a set of healthy images $X_{\text{normal}} \in \mathbb{R}^{x,y,z}$ and learn to project it to and recover it from a lower dimensional distribution $\mathbf{Z} \in \mathbb{R}^{K}$, as can be seen in Figure 2.10 [5].There are different approaches based in this principle:

- Autoencoders: the model is trained to compress and reconstruct healthy anatomy by minimizing a reconstruction loss L. It does so by training an encoder and a decoder architecture that learns how to compress and decompress the input information. There have been different approaches based on autoencoders for anomaly detection on CT as the ones proposed in [4, 71]. Lately, the use of autoencoders with transformers has shown promising results [62].

- Latent variable models: which enforce the latent representation of the input to be similar to the one from the reconstructed output by adding regularization on the manifold structure. One of the most known architectures for this group is variational auto-encoderss (VAEs) [39] where the encoder and decoder learn to parametrize the latent distribution of the data. VAEs have been used for anomaly detection in CT images in the work proposed in [4, 58] and also in other medical imaging modalities [15, 111].

- Generative models: The goal of a generative model is to study a collection of training examples and learn the probability distribution that generated them. generative adversarial network (GAN)s[31] were pioneering in their use for anomaly detection by computing the discrepancy between the input and the image restored by the GAN. GANs are trained by simultaneously updating the discriminator so that it discriminates between samples from the data-generating distribution from

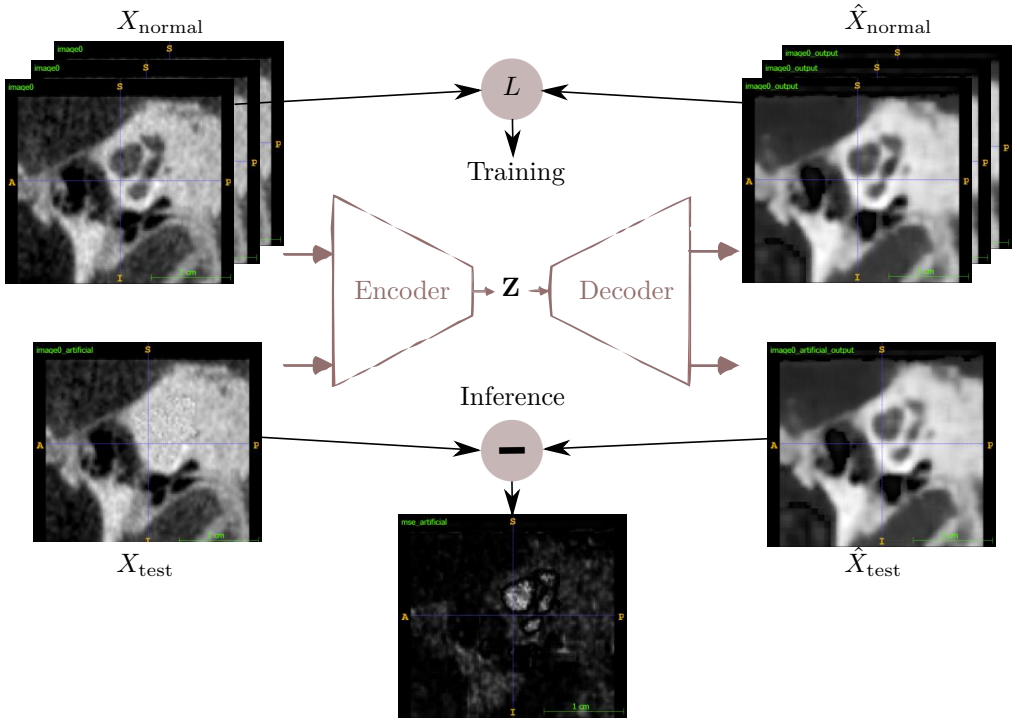


Figure 2.10: Autoencoder-based anomaly detection approach where in training the model learns how to encode and decode healthy data by using a reconstruction loss function L that penalizes differences between input and output; anomalies are detected during inference based on the reconstruction error between input and output.

those of the generative distribution. However, due to their instability in training, they are a less common option, but some approaches, as the one presented in [1, 79], have tried to overcome this issue in CT medical images, it has similarly been done in other medical imaging modalities [74, 6, 73]. The recent advance in the field of generative artificial intelligence (AI) has promising results for anomaly detection in medical imaging as the reversed autoencoders presented in [7] or the use of diffusion models as in the work introduced in [61].

2.2.3 Shape classification

Multiple medical image analysis problems involve the classification of a 3D shape that represents the topology of the anatomical structure of interest. In this section, we introduce how these shapes can be represented, which unsupervised methods are used to classify them, and how graph attention networks can be used for this task and exploit their explainability.

2.2.3.1 Shape representation

The explicit representation of 3D shapes takes different forms. The most common representation of 3D images is using voxels in an Euclidian grid where the intensity value of each of the voxels represents the attenuation of radiation corresponding to the tissue at that location. To analyze a specific anatomical structure,labelmaps are commonly used. Labelmaps can be derived from a manual annotation made by an expert or from a model trained to segment a specific structure or multiple structures. Great advancements have been made to build robust and reliable architectures for 3D image segmentation where most of these are based on CNNs.

The labelmaps can be represented as the voxel volume where the corresponding labels are stored and therefore the physical coordinates and dimensions of the structure of interest are preserved. However, for different visualization tasks, 3D surface meshes are commonly used as they allow for better 3D visualization. When 3D meshes are generated from the labelmap usually some algorithm, like marching cubes [48], flying edges [75] or surface nets [24], is used and it is possible to apply a smoothing transformation that allows the visualization of the 3D surface without the voxelized appearance derived from the limited image resolution.

A mesh is a collection of vertices, edges, and faces that define the shape of a 3D object. There are several types of meshes, such as triangulated-, quadrilateral-, tetrahedral- or hexahedral meshes, that are commonly used in 3D modeling and analysis. Triangle meshes, where the faces are composed of triangles, are the most prevalent due to their simplicity and ease of rendering. Graph- and mesh-based methods have become increasingly important in the analysis of 3D shapes. These approaches utilize the inherent structure of meshes and graphs to perform various tasks, such as segmentation, classification, and feature extraction. Graph-based methods represent the 3D shape as a graph, where vertices correspond to mesh vertices and edges represent the connections between them. This representation allows for the application of advanced algorithms from graph theory and network analysis, facilitating the capture of complex topological and geometric properties. Mesh-based methods, on the other hand, directly operate on the mesh structure, often using techniques like mesh smoothing, simplification, and parameterization to enhance the analysis and visualization of 3D shapes. However, meshes generated from segmentations do not usually present topological correspondence and contain an arbitrary number of edges and vertices, which makes using graph or mesh-based architectures for classification tasks much more challenging.

Another way of representing 3D shapes are point clouds which do not present some of the challenges that are faced with the meshes. When describing a 3D shape with a point cloud, a collection of points with x,y, and z coordinates covering the surface are sampled. Point clouds entail certain topological ambiguity given the lack of connectivity information but allow the representation of highly detailed surfaces. The data structure of point clouds is simple, allowing the coordinates to be directly fed into neural networks for processing. Point clouds have therefore become a very popular choice for object classification. The first DL framework to directly handle 3D point clouds was PointNet [63]. Some of the benefits of this architecture are its invariance to rotations, translations, and point permutations but still combining local and global information making it a very suitable option for classification, part segmentation and scene semantic parsing tasks. PointNet is highly efficient and effective and it was later followed by PointNet++ [64] which captures better the local structures derived from the natural metric space of points by combining features from multiple scales with hierarchy.

2.2.3.2 Unsupervised classification

Unsupervised classification in the context of 3D shape analysis involves categorizing 3D shapes without relying on labeled training data. Instead, these methods leverage intrinsic properties of the data, such as geometric features or statistical patterns, to identify and group similar structures. This approach is particularly valuable in scenarios where labeled datasets are scarce or expensive to obtain. Given a certain encoding architecture, the most common approach for unsupervised classification is building a latent space (a lower dimensionality subspace) that is structured and representative of the features of the input data.

Recently, there has been increased interest in unsupervised learning for feature representation, which focuses on extracting meaningful features from unlabeled datasets. This approach is more challenging than supervised learning due to the absence of direct label guidance. For a fully unsupervised classifier based on a structured latent space, it is important to add constraints in the latent space. There are two ways to train a reliable feature extractor, with a loss function that optimizes the inner- and inter-cluster distances with pseudo labels or with an extra task that can help train the feature extractor [67]. A schematic representation of these two approaches is shown in Figure 2.11. Some common extra tasks, also refer to as “pretext” tasks, for 3D images include self-reconstruction, context prediction, pseudo-label classification, and feature contrast.

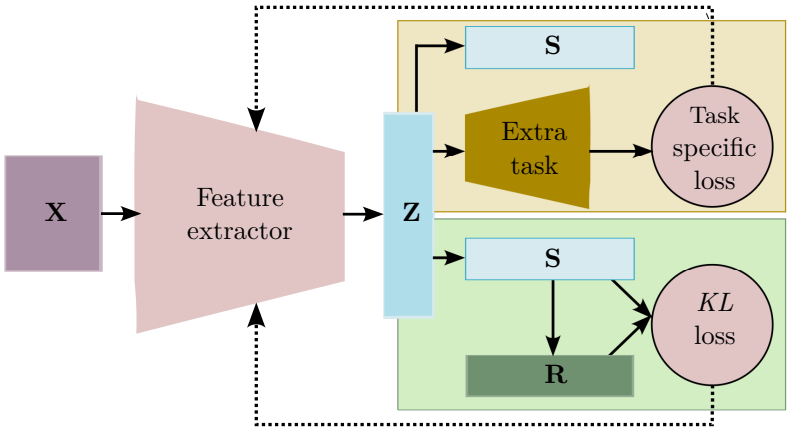


Figure 2.11: Two different approaches for training a feature extractor for unsupervised classification based on clustering of the latent space **Z** where the predicted results are denoted as **S**. The first approach is with an extra task network as shown in the yellow rectangle. The second approach, in the green rectangle, is based on fine-tuning the cluster assignments where **R** is just an adjustment of **S**.

Typical extra tasks for feature extraction in the literature are based on deep autoencoders (like PARTY [60], DSCDAE [103] and AGMDC [92]) or variational autoencoders (like S3VDC [13] or DSVAE [101]), also there is work within adversarial learning (like ADEC [56] or IMDGC [102]) and with deep neural networks (like JULE [100] and DCCM [97]) and, finally, based on graph neural networks (like AGCHK [110] or SDCN [10]).

In the medical image field, as well as in others, domain adaptation is a major issue for these methods that often struggle to extract features that are not domain-specific and that generalize well to a slightly different domain like it can be images coming from a different type of scan or acquisition protocol. There have been efforts towards overcoming these challenges where unsupervised domain adaptation is usually the adopted approach given the lack of domain labels or paired data availability [84].

2.2.3.3 Graph attention network (GAT) and explainability

The importance of moving beyond mere classification in the medical context cannot be overstated. Medical decisions impact patient outcomes directly, and a simple classification result often lacks the necessary depth and nuance required for critical decision-making. For instance, knowing that a tumor is classified as malignant is insufficient without understanding the underlying factors contributing to this diagnosis. This deeper insight is crucial for selecting appropriate treatments, predicting patient outcomes, and providing personalized care. Hence, incorporating methods that enhance the interpretability and explainability of models is essential to ensure that medical practitioners can trust and effectively utilize AI systems in their diagnostic and therapeutic processes.

GATs [86] are a type of neural network architecture specifically designed to operate on graph-structured data. Unlike traditional neural networks that work on fixed-size vectors, GATs leverage the relationships and connections between data points by using attention mechanisms to weigh the importance of each node's neighbors differently. GATs use multi-head attention [85] approach which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension which allows for attending to parts of the sequence differently. This approach allows GATs to dynamically focus on the most relevant parts of the graph, making them particularly powerful for tasks involving complex relational data. GATs assign different weights to nodes based on their influences during feature aggregation. The original GAT [86], inspired various adaptations like C-GAT [91], GATv2 [11], and SuperGAT[38], which employ different attention strategies. GATs can be classified as intra-layer or inter-layer. Intra-layer GATs use attention to weight nodes within local neighborhoods and dynamically update node representations.

Inter-layer GATs use attention for feature combination, selecting features from different levels, channels, views, or time slices [78]. One of the benefits of using these attention mechanisms is their potential for using post hoc explanation techniques.

Explainability in GATs is a vital aspect that enhances the trust, transparency, and usability of these models. While several methods exist to provide insights into how GATs make decisions, challenges remain in terms of scalability, robustness, and human-centric explanations [105].

There are different methods for achieving explainability while using GATs. One of the straightforward methods is to visualize the attention weights assigned to edges in the graph. These weights indicate the importance of neighboring nodes when aggregating information but can sometimes be noisy, especially in complex graphs [41]. Another approach is using node feature importance which assesses the contribution of individual node features to the final prediction using techniques like gradient-based methods (e.g., Integrated Gradients [80]) or perturbation-based methods (e.g., SHAP [50]). These methods can be computationally intensive and may require multiple forward passes through the network.

Using subgraphs by extracting the subgraphs that are most influential for a given prediction using algorithms like GraphLIME [35] or GNNExplainer [104] to identify crucial substructures by perturbing parts of the graph and observing the impact on the model's output. However, extracting subgraphs that faithfully represent the model's reasoning, while still being comprehensible to humans, can be challenging. In a similar fashion, counterfactual explanations are often used where explanations are generated by identifying minimal changes to the input graph that would alter the model's prediction [90]. These techniques involve modifying node features or edges and observing the effects on the model's output, often framed as an optimization problem but still present the same challenges as subgraph extraction.

Finally, there are global explanations methods that go beyond instance-level explanations, these consider understanding the global behavior of GATs to be crucial. This involves identifying common patterns or motifs the model relies on [105]. These approaches often rely on clustering attention patterns across multiple instances or using rule-based systems to summarize model behavior. Obtaining the right balance between complexity and interpretability in these methods still remains a challenge.

CHAPTER 3

Data

3.1 Inner ear datasets

3.1.1 CI dataset

The CI dataset utilized in this thesis is composed of preoperative CT scans of CI surgery patients. It includes 119 CT scans focusing on the inner ear of various patients. These scans, cropped to a region-of-interest (ROI) measuring 32.1^3 mm^3 and centered in the cochlear structure, share a similar orientation and come from diverse and unknown CT imaging equipment. Consequently, this dataset accurately represents the typical anatomical features found in patients and serves as a realistic sample of the imaging data available to physicians for diagnosis and intervention planning of CI therapy.

A sample from the dataset is shown in Figure 3.1. Significant variations in tissue contrast and image quality are evident, primarily due to the differing origins of the imaging equipment, as illustrated by the diverse samples in Figure 3.1. While the resolution of some CT images in this dataset may not always be sufficient to clearly delineate the neural structures of interest throughout their anatomical path, they nonetheless represent the standard clinical images used in this field. Additionally, the dataset includes images of both the right and left inner ear. Given the high degree of symmetry in this anatomical area [66], all images from the right inner ear have been mirrored to reduce variability across the dataset.

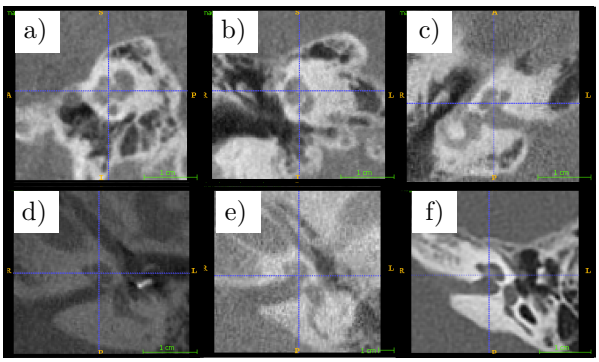


Figure 3.1: Top row: Normal CT of a left ear from the CI dataset a) axial plane b) sagittal plane c) coronal plane. Bottom row: Axial plane of different CT scans of the right ear from the CI dataset to show its variability d) sample with an artifact e) sample with more noise f) sample with higher contrast. The blue cross is approximately located in the center of the cochlea for all the samples.

3.1.1.1 Landmark annotation

The CI dataset has been annotated with a set of 7 landmarks designed to characterize the FN and the CN in the proximity of the cochlear structure. Deciding the anatomical position of the landmarks is a key process. The landmark must be as unique as possible within the structure so it can be easily differentiated from other points along the structure. This gives more robustness to the manual annotation which also influences the robustness of the model as its training process is based on these annotations.

Given the dimensions and lack of contrast of the nerves in these CT images, it is not possible to locate very specific features of the nerves. Therefore, after analyzing both the anatomical structure and its visualization in the CT scans, 7 landmarks were designed accordingly.

Two of the seven landmarks are used for the CN and the other five for the FN characterization. The distribution of the landmarks can be seen in Figure 3.2 and they can be described as:

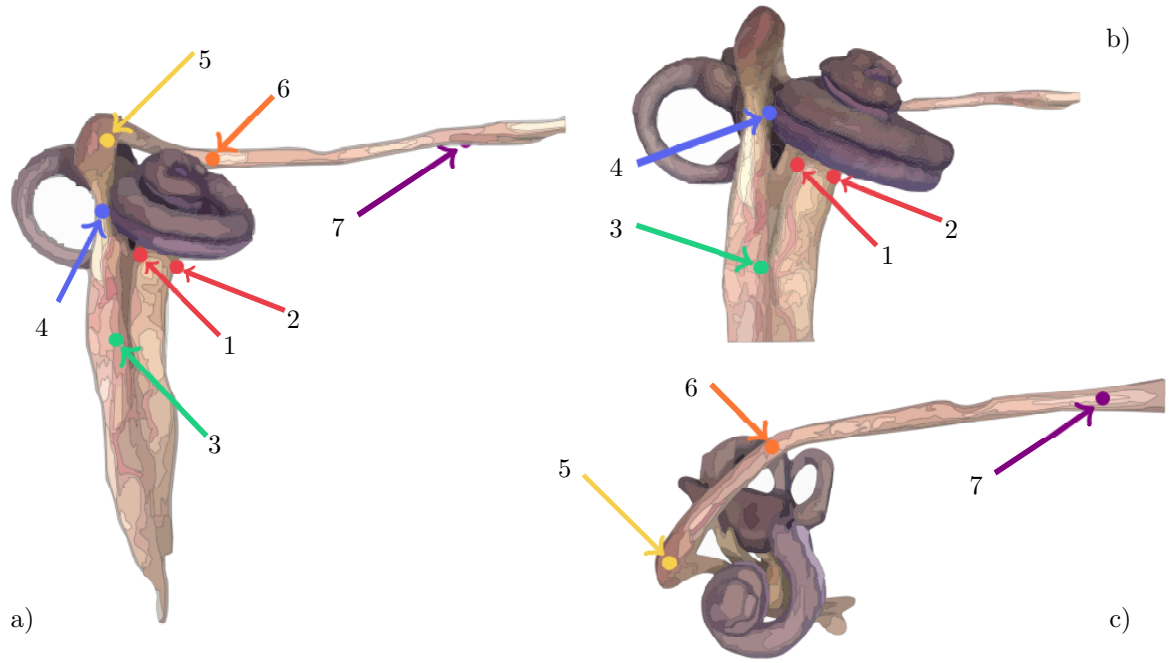


Figure 3.2: Landmarks' location within the FN and CN. a) An overview of the 7 landmarks b) Close-up of the 4 initial landmarks and the labyrinthine segment c) Close-up of the 5-7 landmarks and the tympanic segment.

- 1 and 2 - define the diameter of the CN from the axial view of the CT. Their estimation is relatively easy due to the good contrast present at this anatomical point as well as the closeness to the very characteristic cochlear structure.
- 3 - identifies the point at which the FN joins the IAC. This landmark also defines the end of our ROI regarding the FN.
- 4 - identifies the closest point between the FN and the cochlea. It is not an easy landmark to annotate due to the great proximity that sometimes occurs between both structures and the resolution of the scan. However, this landmark is vital as it will prevent any possible leaks between both structures and locates a relevant point regarding the FNS due to proximity with the electrode array described in section 2.1.1.2.
- 5 - is placed in the geniculate ganglion, the landmark has been placed close to the bifurcation and we expect that it will help to characterize the great curvature that the neural structure presents at this point. This region usually lacks contrast with the surrounding tissue, especially in the part that is further away from the bifurcation.
- 6 - defines the direction after the genu. It is placed closely after the nerve passes the semicircular canals. This is a complicated region because the width and intensity levels of the semicircular canals in the CT are, on many occasions, rather close to the FN and it can be confusing for the annotator.

- 7 - is placed close to the end of the CT scan in the area where the FN is straight and relatively easy to locate. However, the variability regarding the longitudinal point of the FN in which this landmark is located is relatively high. There is no significant feature regarding this landmark besides the proximity to the end of the CT scan. It is however not a crucial landmark and mostly defines the end of our ROI within the FN.

The chosen tool for placing the landmarks is *ITK-snap* [106] visualization software. During the annotation process, 119 CT scans have been processed. To get a better overview Figure 3.3 shows the landmarks, described in the previous section, placed in a randomly selected CT from the data set.

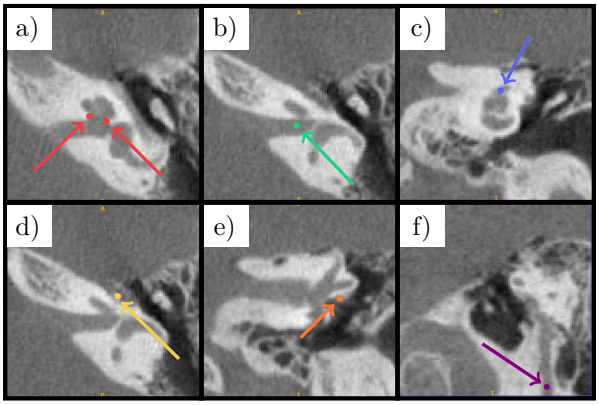


Figure 3.3: Landmarks' location in a sample CT scan. a) 1 and 2 landmarks in axial view, b) Landmark 3 in axial view, c) Landmark 4 in coronal view, d) Landmark 5 in axial view, e) Landmark 6 in coronal view, f) Landmark 7 in sagittal view. For each landmark, the most representative visualization plane has been chosen.

3.1.1.2 Artificial malformations

The artificial dataset consists of the artificially generated pairs for the 119 clinical CT scans of the CI dataset which consists of patients with normal inner ear anatomy described in section 3.1.1. We synthetically generated abnormal inner ear CT scans from the original images by removing the cochlea (simulating cochlear aplasia), thus generating corresponding pairs of normal and abnormal CT scans with the same surrounding structures. The cochlea was segmented using ITK-SNAP software [106] and then replaced by Gaussian noise with mean and standard deviation estimated from the intensities of the tissue surrounding the segmentation. An example of the inpainting process is shown in Figure 3.4.

3.1.2 Congenital malformation dataset

This clinical dataset consists of 122 CT scans of inner ears that present different types of IEMs as well as 300 anatomically normal CT scans from heterogeneous sources. The ROI extraction for this dataset was done using the methodology described in Contribution *B* using anatomical points of interest that were not involved in the anatomy of interest to allow for a standardized and robust image orientation regardless of the appearance of the inner ear region. A greater ROI of $80^3 mm^3$ was selected for this dataset in order to contain all the anatomical points of interest for CI therapy.

For the 122CT scans of single inner ear anatomy with some congenital malformation, the imaging equipment and protocol used for the acquisitions are different and unknown therefore the quality and appearance of the images varies wildly. The anatomically normal CT images have been collected from multiple centers and also present different quality levels and overall appearances to keep both classes as heterogeneous as possible. The congenital malformations included in this dataset were clinically explained in section 2.1.1.3. However, IEMs are not equally common, our dataset encapsulates the

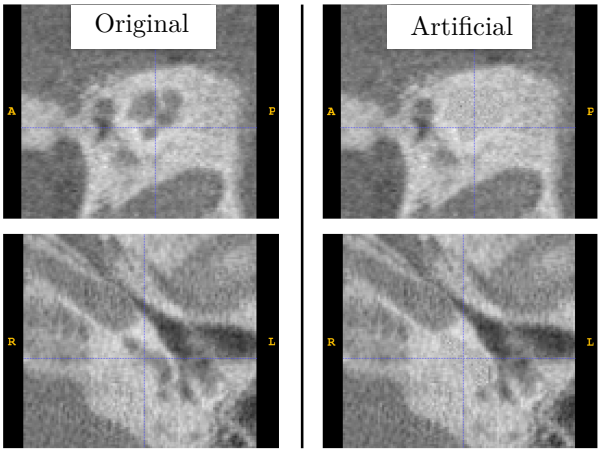


Figure 3.4: Sample CT scan from artificial malformations, original slice of CT image on the left and same slice of artificially generated malformation on the right.

frequency of these anatomical malformations and therefore presents quite an unbalanced distribution which can be observed in Figure 3.5.

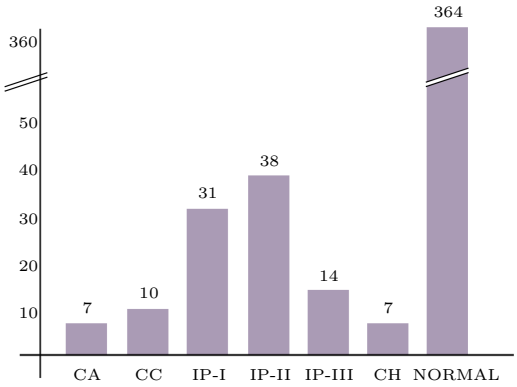


Figure 3.5: Distribution of cases among the different types of IEMs: cochlear aplasia (CA), common cavity (CC) incomplete partitioning type I, II, and III (IP-I, IP-II, IP-III), cochlear hypoplasia (CH), and normal.

3.1.2.1 Landmark annotation

For this dataset, as shown in Figure 3.6, twelve anatomically relevant landmarks were carefully designed and annotated in a randomly selected subset of 160 CT scans of anatomically normal cases in collaboration with our clinical partner, an ENT surgeon specialized in CI therapy in abnormal anatomies.

Annotating landmarks in an abnormal anatomy is sometimes not possible. As previously described, landmarks need to be well defined both conceptually but also be as unique as possible anatomical points in the image. The congenital malformations are a broad spectrum and different parts of the anatomy can be severely affected by the malformation or sometimes, in the more extreme cases, not even present as described in section 2.1.1.3. Therefore, the landmarks that were defined for this research avenue were annotated only in the anatomically normal images of the dataset. There is a total of 12 landmarks which consist of:

- 1 and 2 - located in the sigmoid sinus and the external acoustic canal respectively. Both landmarks are located at the closest point between both anatomical structures.
- 3 - located in the jugular bulb's most proximal point to the round window.
- 4 and 5 - located in the carotid artery and the basal turn of the cochlea. Both landmarks are located in the most proximal point between both structures.
- 6 and 7 - anterior and posterior edges of the round window, which is the anatomical opening of the cochlea through which the electrode is inserted.
- 8 and 9 - anterior and posterior crus of staples, which is the third ossicle located the closest to the cochlea.
- 10 - located in the short process of incus, which is the second ossicle located between the maleus and the staples.
- 11 - located in the pyramidal process.
- 12 - located in the cochleariform process.

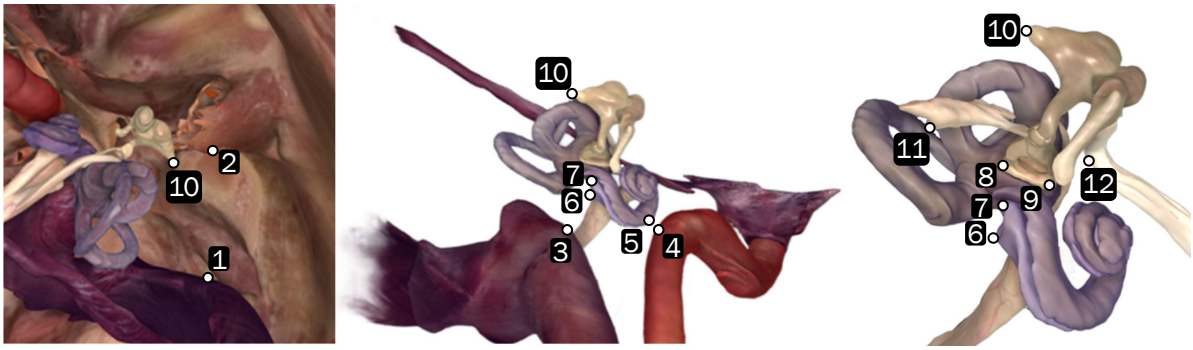


Figure 3.6: Set of landmarks annotated in the anatomical normal CT images of the congenital malformation dataset. 1 - Sigmoid Sinus . 2 - External Acoustic Canal 3 - Jugular Bulb . 4 - Carotid Artery. 5 - Basal Turn . 6,7 - Anterior and posterior edges of the round window. 8,9 - Anterior and posterior crus of staples. 10 - Short Process of Incus. 11 - Pyramidal Process. 12- Cochleariform Process.

3.1.3 ROI dataset

This dataset consists of 140 full-head CT scans from the CQ500 dataset [16] which corresponds to 102 different patients. The dataset has been made publicly available by the Centre for Advanced Research in Imaging, Neurosciences and Genomics of New Delhi. The CT scans come from several radiology centers in New Delhi and are collected using seven different CT scanners. Furthermore, the images were performed to detect the presence or absence of an intracranial hemorrhage and we can observe that the resolution is poorer for inner ear analysis than expected.

3.1.3.1 Landmark annotation

All used scans are resampled to have the isotropic voxel spacing 0.5 mm. The image dimensions vary significantly within our dataset. All scans are manually labeled with the chosen landmarks shown in Figure 3.7 and all the annotations were made publicly available as shown in Contribution *B*.

Choosing relevant landmarks is necessary to characterize the inner ear orientation. The landmarks must be uniquely defined within their structure, so they are easily differentiated from other anatomical

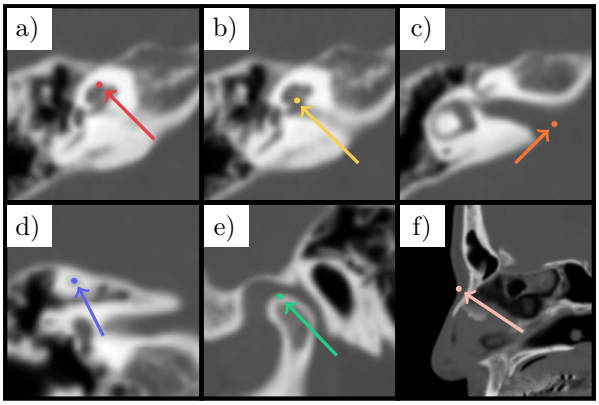


Figure 3.7: Landmarks 1-5 and 11 are shown on CT scan a) landmark 1, b) landmark 2, c) landmark 3, d)landmark 4, e) landmark 5, f) landmark 11. Landmarks 1-5 are for the right inner ear ROI, but the corresponding landmarks 6-10 are placed at the same anatomical points on the left side of the head.

points nearby. Eleven landmarks are chosen in total, five assigned for each inner ear and one outside this region. The five landmarks for each ROI are the same anatomical points but located on their respective side of the CT scan. The landmarks can be described as:

- 1 and 6 - located in the cochlear apex of the left and right side respectively.
- 2 and 7 - located in the cochlea nerve, in the midpoint below the base of the cochlea in the left and right side respectively.
- 3 and 8 - located at the end of the IAC in its center left and right side respectively.
- 4 and 9 - located in the superior semi-circular canal peak of both the left and right side respectively.
- 5 and 10 - located at the top of the condyle in the mandible, where the temporomandibular joint starts for both the left and right side respectively.
- 11 - located at the nasion which is the point in the skull at which the suture between the two nasal bones meets the suture between these and the frontal bone.

3.2 Heart dataset

3.2.1 PDL dataset

The dataset utilized in this study contains a cohort of 125 patients who have undergone LAAO procedures. For each patient, we collected both pre-operative and post-operative CT images. The pre-operative CT images were used to extract anatomical graphs characterizing each patient's specific anatomical structures, and the post-operative 3D images to evaluate the presence of PDL following LAAO.

The preprocedural scan consists of two volume scans post contrast injection in the arterial and venous phases. The scans are ECG-gated, the first CT, in the arterial phase, is triggered to start when the Hounsfield units in the LV reach 180. The second one, in the venous phase, is triggered 20 seconds later. Sometimes due to a slower blood flow in the LAA, there are partial filling defects in this region, to differentiate between contrast gaps generated by a slower blood flow or the presence of a blood clot in the LAA the CT scan in the venous phase must be analyzed.

When analyzing the post-procedure CT images for leak detection the different types of leaks observed were further classified based on their nature. The population distribution and visual examples of each

leak type are illustrated in Figure 3.8. This figure provides an example CT image for each type of leakage from our study. Blue arrows in the figure highlight regions where the gaps associated with PDLs can be observed.

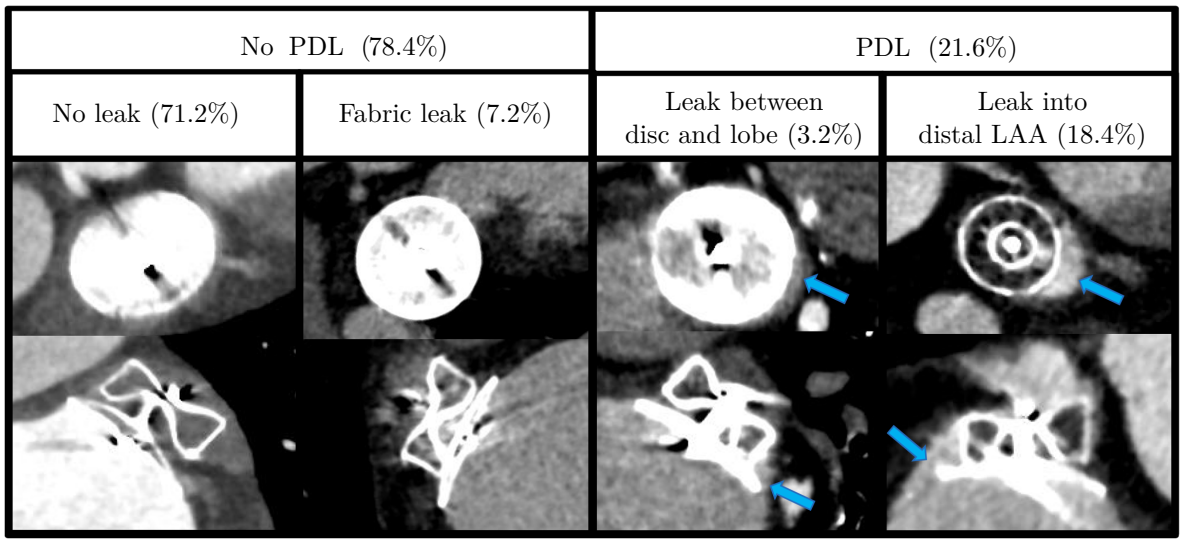


Figure 3.8: Four classes defined based on the mechanisms of leaks after LAAO with a disc-and-lobe occluder. The prevalence of each class over the 125 subjects used for this study is also shown. The implanted device is shown in two views for each example CT image. Blue arrows indicate the regions where the gap associated with the PDL can be observed.

All patients were treated by a consistent medical team, using three distinct types of occlusion devices. The devices include two disc-and-lobe occluders: the Amplatzer Amulet™ (used in 32.8% of cases) and the Omega™ (used in 36.8% of cases), and one plug-in type device: the Watchman™ (used in 30.4% of cases). Leaks were categorized into PDL and non-PDL cases. A PDL is identified by a discernible gap between the device and the anatomical structure, even if the gap is only in the disc part while the lobe is adequately sealed. Non-PDL cases include scenarios where there is either no patency or patency without a visible gap; these typically correspond to fabric leaks or micro leaks expected to seal over time [72]. In some instances, the CT scans were obtained before the process of endothelialization was complete, thus these are not considered PDLs in our dataset. It is important to note that due to the nature of the occluders, only disc-and-lobe devices can exhibit leaks between the disc and the lobe, whereas fabric leaks or micro leaks can occur with all types of devices.

3.2.1.1 Data processing

For segmentation purposes, we employed TotalSegmentator (TS) [95] to obtain labelmaps of the LA, inferior vena cava (IVC), and RA from the pre-operative CT images, an example is shown in Figure 3.9. However, given the highly variable morphology of the LAA, TS was only able to locate it without providing a precise segmentation of this structure. To address this, we used the NUDF method proposed in [77], which utilizes a neural unsigned distance field to generate significantly more precise and detailed segmentations of the LAA. This method enhances the accuracy and reliability of the anatomical analysis in our study. The post-operative CT images were meticulously analyzed by two expert annotators. Their objective was to determine the presence of leaks in the LAA, which was identified by observing whether the contrast agent permeated the distal region of the LAA and whether there was a visible gap between the device and the anatomy.

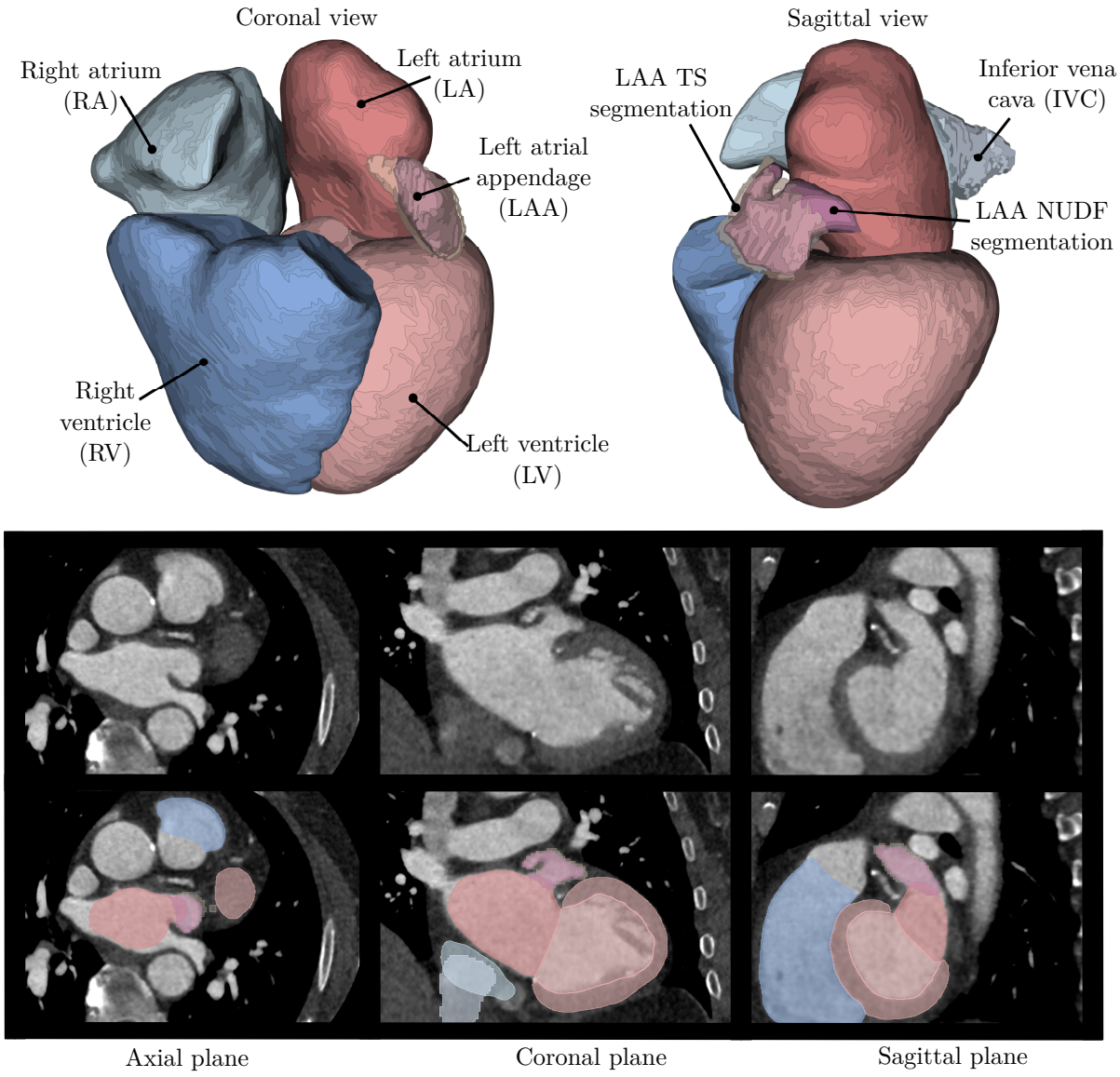


Figure 3.9: Top: 3D visualization of the heart chambers, the LAA and IVC from TS and the refined segmentation of the LAA obtained with the NUDF method. **Bottom:** Sample CT image from the PDL dataset including the overlay with the same labels as in the 3D representation in the top figure.

CHAPTER 4

Thesis contributions

This chapter details the contributions of this PhD project, highlighting their significance, discussing the proposed solutions, and connecting them to each other and the broader context. For in-depth information on each contribution, please refer to the publications in Appendices A-G.

We have separated the contributions into three sections. The first one is focused on applications of DRL for landmark location in clinical CT images; the second constitutes the work on anomaly detection for inner ear anatomy and, in the final section, we present the contributions that focus on the utilization of 3D meshes for medical image analysis.

4.1 DRL for landmark detection

Landmarks play a crucial role not only in medical image analysis but also in various other applications within the field of computer vision. The challenge of defining and extracting reliable landmarks, that can be used to characterize the anatomy or as seeds for subsequent processing tasks, has been a significant area of research in the medical imaging community for many years.

One major advantage of using landmarks is the efficiency it brings when working with unlabeled datasets. Annotating landmarks is considerably faster compared to generating pixel-wise annotations, particularly in the case of 3D medical images, where slice-by-slice annotation is extremely time-consuming and costly.

DRL offers a promising, human-like approach to landmark localization, making it an area worth exploring to understand its benefits and limitations. Building on the C-MARL framework [43] for landmark detection, we have developed methods to better characterize challenging anatomical structures and identify regions of interest in 3D images.

4.1.1 Nerve characterization

Soft tissues, such as nerves, are difficult to distinguish in CT images. However, being able to characterize them in this imaging modality presents a clear advantage for preoperative CI therapy planning. The facial and cochlear nerves, which we focus on in Contribution *A*, are particularly challenging to characterize along their entire path, especially in the region near the cochlear structure as can be seen in Figure 4.1. However, there are specific points where the nerves can be more easily outlined due to their anatomical shape or the contrast with surrounding tissues. Identifying these key spots in the neural structures is essential for accurately characterizing the nerves. While fully annotating the entire nerve is extremely challenging, and in some cases, nearly impossible, the landmarks we have designed and presented in Paper *A* are clearly defined for the annotator, providing a practical solution for nerve characterization.

In the work presented in Contribution *A*, we use the located landmarks as seeds for the characterization algorithms. We evaluate the final landmark predictions not only numerically but we also analyze their qualitative performance, which, for the final objective of our characterization pipeline, is more relevant. The precision required for landmark placement is determined by the subsequent characterization algorithm. Therefore, the quality of landmark placement—specifically, ensuring that it is anatomically correct—is more important than the actual error in millimeters, especially when characterizing tubular structures with landmarks.

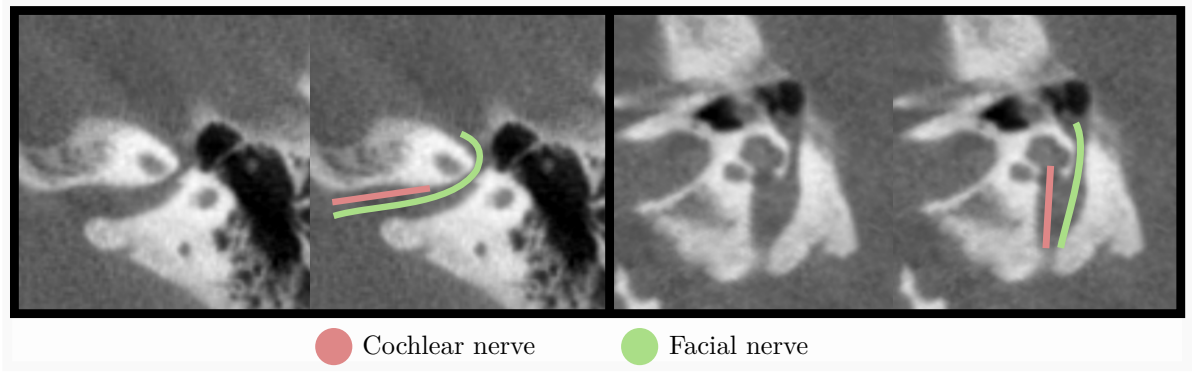


Figure 4.1: Two CT planes showing the estimated path of the cochlear and the FN in the inner ear.

The pipeline presented in Contribution *A* consists of a C-MARL [43] DRL architecture for automatic landmark detection on the CT images. We trained 3 agents per landmark which allows us to analyze the variability and prevent odd behavior by having 3 tentative locations for each specific landmark. The DRL model is followed by specific characterization methods that use the estimated locations as seeds. For the FN we develop a path selection algorithm that uses Dijkstra's shortest path algorithm [23] and the gradient along the path's intensity profile in different path candidates to select an optimal neural path. For the CN, clinicians usually only measure the diameter which is defined by two of the designed landmarks (1, 2 - see Figures 3.2 and 3.3). However, we use these landmarks to extract the crossectional view of the nerve and use the Chan-Vese approach [14] for active contours without edge to segment it accurately. We consider the crossectional area of the nerve to be a much more representative measurement of this anatomy.

The result of this work is a full processing pipeline that automatically characterizes both neural structures in the cochlear region. We developed an advanced tool for CI surgical planning based on deep learning and computer vision algorithms to support clinicians in the routine assessment of FN anatomy, related risk assessment of FN stimulation, and the evaluation of CN health. Furthermore, this work was the initial step towards further exploiting this DRL model and the implicit and explicit information encoded in this model which will be discussed further in section 4.2.

4.1.2 Robust region of interest extraction

The analysis of medical imaging data often involves handling a wide range of variability, both in the images themselves and in their clinical contexts. This variability stems from differences in imaging equipment, acquisition protocols, and patient positioning, all of which contribute to the complexity of developing robust automated processing techniques. We aim to address some of these challenges by providing a reliable and robust way to extract precise ROIs.

Most of the CT images used in Papers *A*, *C*, *D*, *E*, and *F* are full-head CT images that have different origins. The heterogeneity of our datasets is beneficial for simulating the actual clinical scenario and the need for robust methods that are not very sensitive to small domain shifts, as can be the processing protocol, the type of scan, or even a specific watermark. There is significant variability in the position of the head within the 3D image, which presents a challenge for the automated processing pipeline. In Paper *B*, we present a method based on the same architecture for landmark localization that we proved successful in Paper *A*, with the communicative multi-agent approach, but now we use it to correct the head orientation and provide a robust ROI extraction of the inner ear for different CT images.

We chose to use a landmark-based approach which we consider will be more robust and could potentially provide more explainability or even help detect abnormalities from an early processing stage. Furthermore being able to design your own landmarks that will be used to characterize the ROI provides a very significant advantage especially in the case of abnormal anatomies, as you do not need the algorithm to use the ROI, that contains the anomaly, but you can rely on other more general

and robust anatomical features. Another advantage of the DRL method that we use is the multi-scale approach in which the initial field of view is bigger but with a worse resolution, and later the agents zoom in on their area of interest, this process is illustrated in Figure 4.2. This is especially useful as in these heterogeneous datasets the field of view of the scan varies significantly and preventing this variation from affecting the extracted ROI is key.

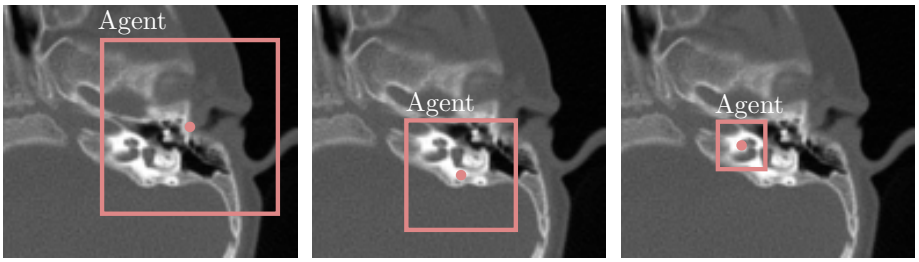


Figure 4.2: Multi-scale approach for landmark detection with hierarchical steps by reducing the step and the ROI size when the agent enters starts oscillating. The episode is terminated when all agents reach their terminal states at the lowest scale.

In Contribution *B* we use eleven landmarks to extract two different ROIs in each head scan, one for each inner ear region. We analyze the robustness of the method by applying a grid of different rotations to the images and comparing performance, with a very positive outcome. Landmarks were localized with an estimated error between 0.78–2.11 mm (on average 1.07 mm). The study presented in Contribution *B* successfully used a DRL framework for landmark detection in full-head CT scans, and utilized it for ROI extraction of the inner ears.

4.2 Anomaly detection

Supervised approaches for anomaly detection have high accuracy but require two things that are sparse in the medical imaging field. The first one is a big and representative datasets that faithfully represent not only the healthy population but also all the possible anomalies as the behavior of supervised models when facing new out-of-the-distribution samples is often unpredictable. The second main problem is that, given that such a dataset exists and is available, it must be labeled by expert annotators which can be expensive, especially for rare pathologies.

We instead tackle the anomaly detection problem with a parametric approach instead of a classification one, to build implementations that move toward more interpretable results. A common approach for anomaly detection is a semi-supervised approach that is trained only in the healthy population. As was described in section 2.2.2, the most common approaches are based on learning from a set of healthy images and how to project it to and recover it from a lower dimensional distribution. Similarly, we train a model for landmark estimation in healthy subjects and then evaluate how to detect an abnormal anatomy from the implicit and explicit information that this model holds.

We have used the proposed DRL models for landmark detection in Contributions *A* and *B* in CT images of the head with great accuracy, specifically in the inner ear anatomy. The model that was used for nerve characterization in Contribution *A* was trained to detect landmarks in the nearby region on the cochlea but not the cochlea itself. Given that the cochlea is the anatomical structure that is more representative of the different types of malformations in this region, we explored how the presence of a malformation would affect the predicted set of landmarks. Could we detect the presence of a malformation by evaluating the configuration of the predicted landmarks? In Contribution *C* we build the proof of concept of how to do so. We generated a synthetically malformed dataset by completely removing the cochlear structure, which simulates cochlea aplasia, illustrated in Figure 3.4. In Paper *C*, we show that building a point distribution model (PDM) that takes into account the overall shape

distribution of the landmarks is more effective than only using the Euclidean distance in the image space (representative of multiple agents' level of agreement). Our approach shows an area under the ROC curve of 0.97 and 96% accuracy in the detection of these synthetically generated abnormal anatomies. The variability is evaluated over multiple inferences, the agents are always randomly initialized within the image, which makes the landmark prediction a stochastic process which we benefit from in this approach to detect malformed anatomies.

Using the predicted configuration of the landmarks in a subspace formed by the PDM of the normative landmarks was a relevant and successful indicator for anomaly detection but we also wanted to investigate whether the model possessed some implicit information about the possible presence of an abnormal anatomy. In Contribution *D*, we introduce a measurement of normality derived from the predicted Q-values of the agents in the last stages before finding the landmark. When the anatomy that the agent is looking at during the final stages does not resemble the normal anatomical configuration of this region, the expected rewards of taking the different actions towards the target present a less uniform distribution. This means that when the anatomy differs from what has been seen during training the model tries to push the agent away and this is shown in the measurement of the variability within the distribution of the Q-values of the action set. A schematic illustration of how this measurement denoted U_{image} and the one derived from the explicit landmark distribution D_{image} are computed is shown in Figure 4.3.

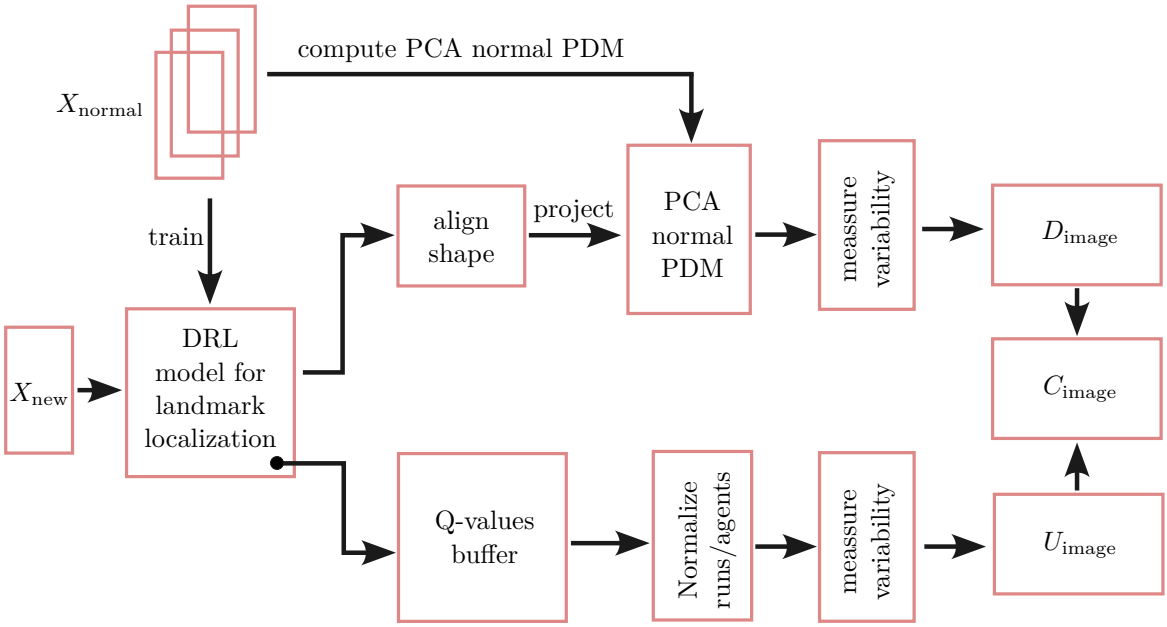


Figure 4.3: Diagram of the DRL approach for anomaly detection based on a model trained for landmark localization, the figure shows the schematic approach to derive the explicit abnormality measurement D_{image} based on the landmark configuration and the implicit abnormality measurement U_{image} derived from the buffer of the Q-values associated to the different agents in the last stages of their search. C_{image} is a weighted combination of U_{image} and D_{image}.

Furthermore, in Contribution *D* we evaluate the two proposed abnormality measurements, not only on the synthetically generated dataset with the set of landmarks introduced in Contribution *A*, but also in a large clinical dataset of congenital IEMs with a specifically designed set of landmarks. The dataset consists of 300 anatomically normal CT scans from heterogeneous sources and 122 CT scans of inner ears that present diverse congenital malformations as the ones described in section 2.1.1.3. For this new dataset, we designed a specific set of 12 landmarks which were picked to be representative not only of the anatomical variations present in the very different types of malformations but also to be

representative of the surgical approach and of the key features that the clinician will evaluate during CI therapy planning. These designed landmarks could potentially give insights into which specific part of the anatomy is influencing the decision and therefore produce some explainability. In Paper *D*, we observed that the DRL model called MARL, in which there is no explicit communication between agents, was more successful in detecting the abnormalities. This is the result of a combination of factors. The explicit communication allowed the agents to end up having some knowledge about their peers' whereabouts which favors that they will end up in a plausible distribution according to the PDM of the normal anatomies. Even if the anatomy is not as expected this communication can potentially also influence the level of hesitation giving the agents false certainty regarding the landmark localization task. All in all, Paper *D* shows that this extracted information that looks at the predicted landmarks over different runs/agents in a subspace defined by the normal anatomies can be complemented with the distribution of the Q-values of the last interactions of the agents providing a reliable anomaly detection measurement for CT images which showed to work well, not only in the synthetic datasetc, but in the very challenging clinical dataset where it achieved an area under the ROC curve of 0.87.

One of the main challenges of the clinical dataset that contains the inner ear CT images of cases with congenital malformations is that they come from different and very heterogeneous image equipment. These cases are rare and only a few ENT specialists can asses them, that is why images from different centers in different countries are added to the same database. Building a model that is reliable and robust to this heterogeneity is a big challenge. In our research, we have used a multi-scale approach, illustrated in figure 4.2 which means that the agent initially has a wider field of view with a poorer resolution, and as it gets closer to the landmark the field of view is reduced and the resolution improved. This means that during the last stages of the search, the agent has access to more local features. One of the benefits of using our proposed approach for anomaly detection is that the DRL agents make decisions based on more local structures and therefore we expected our method to be more robust to changes in the overall image appearance. We test our hypothesis in Paper *E* where we closely benchmarked our method against a 3D convolutional autoencoder (CAE) with an asymmetric architecture developed for anomaly detection in emergency head CT volumes proposed in [71].

CAEs are an unsupervised learning algorithm that learns an identity mapping of the input by minimizing the loss function between the input and its reconstructed output. They use deep convolutional layers to perform the dimensionality reduction. The local connectivity of convolutional layers enables the CAE to extract local and hierarchical features capturing the global feature of the input by combining the local features. The underlying principle is to use the difference between the original image and the one produced by the generative branch of the model to estimate the probability of the given sample being an anomaly.

We compare their performance in both artificially generated anomalies and clinical images of congenital IEMs. The DRL-based method outperforms the 3D-CAE mostly due to a better adaptation to the heterogeneity of the clinical datasets. The image quality affects more the direct output in the 3D CAE than in the DRL approach, on top of that, for the CAE it also affects the error value (which is directly related to the anomaly measurement) and its distribution. In this case, we used the patch-based MSE (A_{image}) which helps alleviate this influence, but we still observed, especially in the synthetic dataset that allows for direct comparison (pairwise), how the overall appearance of the image directly affects the error value and distribution sometimes more than the artificial malformation itself. This is partly also given the smooth appearance of the autoencoder output that can be observed in Figure 4.4 which is a common quality of these types of approaches that sometimes are used for denoising data.

In Figure 4.5 we can observe the ROC curves for both our DRL approach and the 3D-CAE approach in the artificial but also the clinical dataset, we observe that the DRL approach presents better ROC curves especially when looking for low false positive rates. When we compute the area under the curve (AuC) the difference in performance between any of the DRL methods and the 3D-CAE is very clear (11, 2% improvement on average).

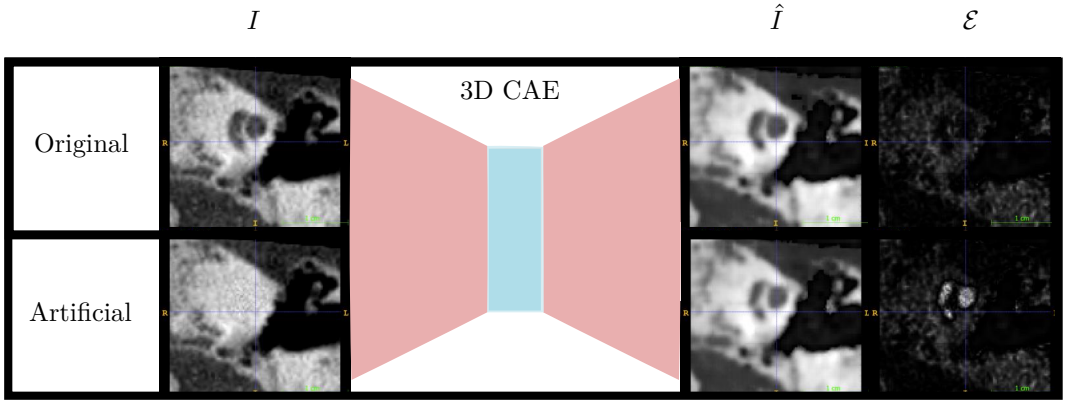


Figure 4.4: 3D-CAE test example from an original CT image from the CI dataset and its corresponding artificial malformation (corrupted version with in-painted cochlear structure). I is the input CT image and $\hat{I}$ is the output of the autoencoder, meaning the reconstructed version of the input image. $\mathcal{E}$ shows the reconstruction error between input and output.

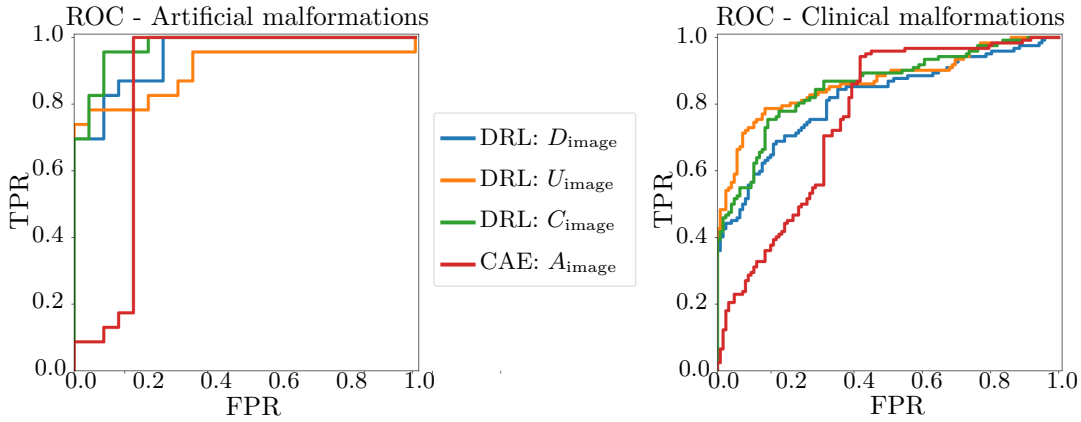


Figure 4.5: Evaluation in the artificial (left) and clinical (right) datasets. ROC curves were computed for each of the methods based on DRL (green, blue, and orange) and the 3D-CAE method (red).

4.3 Analysis of 3D shapes from medical images

In Contributions F and G, we focus on working with 3D shapes. These shapes are extracted from the CT images with different automatic methods that generate the segmentations and corresponding 3D meshes of the structures of interest. When we work with shapes we lose some of the image information as the intensity values and surrounding structures but it allows us to focus on the shape analysis and mitigate some of the challenges of heterogeneous datasets.

4.3.1 Unsupervised classification

Detecting the presence of a congenital malformation in the inner ear with the methods described in Papers C, D, and E is an important task for screening and assisting doctors in their daily practice. However, given the complexity of evaluating this anatomy and generating a diagnosis of the type of malformation, we explored different options to provide relevant information about the type of congenital malformation present in a specific patient. Thus, when a case is detected as abnormal, we can further categorize it within the broad spectrum of congenital malformations.

We faced two main challenges regarding the ability to classify the types of malformations given the dataset available for this purpose:

- The first challenge is the heterogeneity of the images' source and acquisition protocols and, therefore, overall appearance. Most approaches for assisted medical image diagnosis are commonly developed using quite homogeneous datasets that are not representative of real-world applications and the diverse clinical settings in which they must operate. To reduce the influence of the image appearance, in our approach we decided to extract the 3D meshes of the structure of interest and work directly with solely this representation.
- The second main challenge is the data distribution which involves a big class unbalance with some classes having a very limited number of samples. This can be observed in Figure 3.5. This limiting factor motivated us to find an unsupervised approach in which we could build a space where the different types of malformations were structurally represented and therefore the classification of the subgroups or cluster creation was possible but without including information about the type of malformation in the training process.

The pipeline presented in Paper *F*, which is shown in Figure 4.6, is designed with a profound comprehension of this data type and the congenital malformations themselves. We have observed that the cochlear structure can be roughly but consistently segmented by a 3D-UNET [70] model trained exclusively on normal cochlear anatomies. We then use these segmentations and adopt an entirely unsupervised approach, meaning the DL model is trained from scratch on these segmentations, and the class labels are not used for training. To map these shapes to an optimal latent space representation, we utilize DeepDiffusion [25]. This algorithm optimizes both the feature extraction and the embeddings produced by the encoder, which results in salient features in a continuous and smooth latent space. To do so, the algorithm uses the latent manifold ranking loss which contains two terms, one that pulls all the augmented versions of the same sample together (L_{fit}) and another that pulls extrinsic features and their neighboring intrinsic features together (L_{smooth}).

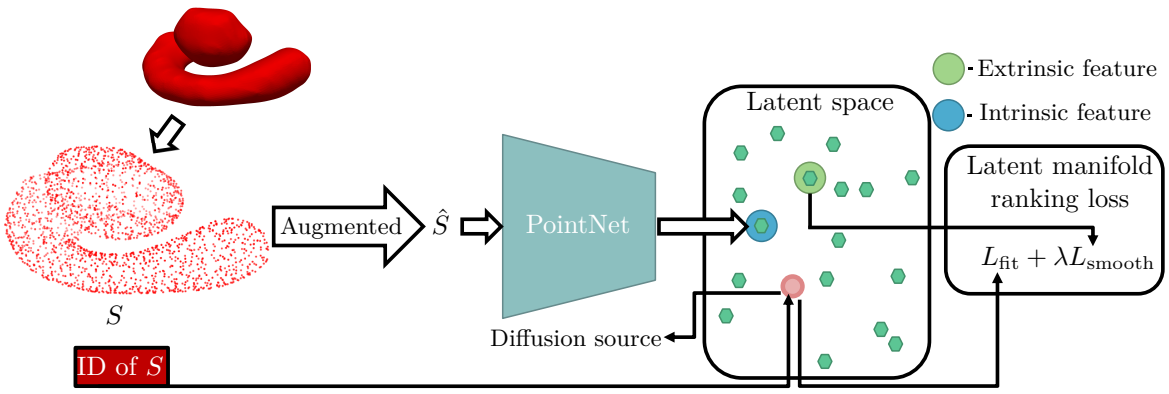


Figure 4.6: Sketch of the unsupervised pipeline used in Contribution *F*. We use DeepDiffusion for latent space representation of the cochlear 3D meshes. The point cloud extracted from the mesh is fed to the PointNet encoder which generates the corresponding latent feature which is optimized by minimizing the latent manifold ranking loss so the encoder and the latent feature manifold are optimized for the comparison of data samples.

One of the difficulties of working with meshes generated from binary segmentations is that they are unstructured graphs and are therefore complex to standardize for a DL approach, in our method we sample 1024 representative points from these very different surfaces and then use PointNet as the encoder part of our model to overcome this challenge. This point-cloud approach allows for each case to be represented by the same number of sampled points and therefore have the same dimensions and structure.

Figure 4.7 shows the latent space distribution of the test samples using UMAP, the representation of the different cases in the latent space shows spatial relation between classes, which is correlated with the anatomical appearance of the different malformations shown in Figure 2.4. In Contribution *F* we showed how using the 3D shape information of the cochlea obtained with a model only trained in normative anatomies is enough to classify the malformations and reduce the influence of the image's appearance, which is crucial in a clinical application setting. Also, the method achieved a mean average precision of 0.77 with a mean ROC AuC of 0.91, the associated ROC curves for each class as well as the pairwise cosine distance matrix between samples of the test set are shown in Figure 4.8. The results indicate the effectiveness of this approach in classifying different inner ear types of congenital malformations.

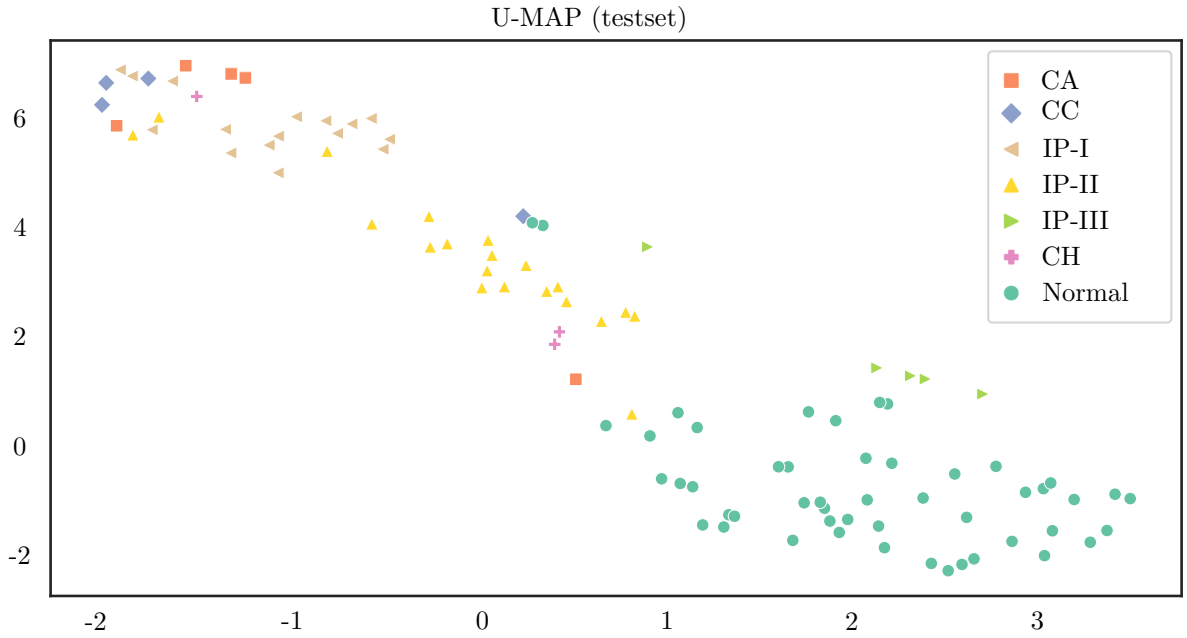


Figure 4.7: Visualization of the U-MAP representation of the test features where it can be observed how the different classes group together and how the anatomical variation is represented as there is a progression from the most abnormal cases towards fully normal cases.

4.3.2 GAT with explainability

In the final contribution of this thesis, Paper *G*, we collaborated very closely with a group of interventional cardiologists in Rigshospitalet. Clinical input is extremely valuable and should be carefully considered whenever we are trying to build AI tools to support clinical practice. Familiarizing with the daily workflow of the clinical practice and the decision process is key to designing useful tools. In the work presented in Contribution *G*, we focus on finding potential image-based biomarkers related to the morphology of the anatomy that could be found for patients who present PDL following LAAO implantation.

PDL is a multifactorial complication, given the many different factors that affect the procedure and the multiple decisions involved in LAAO. Device selection, expertise in device placement, and LAA anatomy are just a few of these factors. Being involved in the procedural planning and observing multiple LAAO procedures in the CathLab gives extremely useful insights. It is known that some anatomies are more difficult for the operator and that they can recognize them from CT images. We wanted to isolate the anatomical morphology and evaluate its impact on PDL risk.

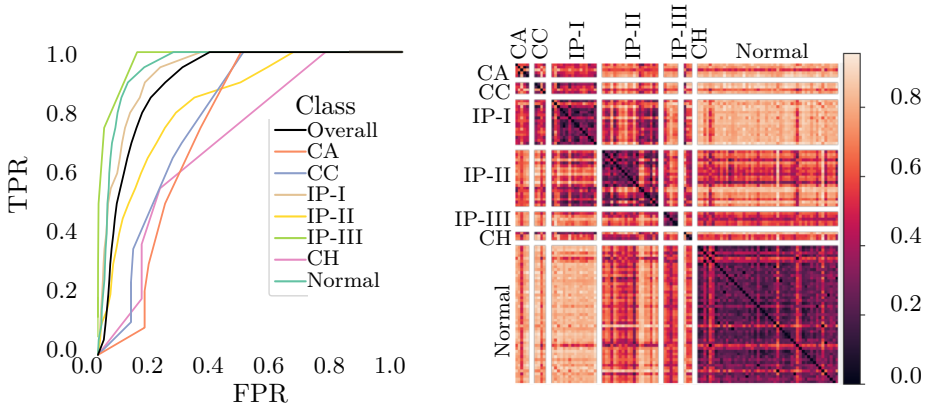


Figure 4.8: Evaluation of unsupervised classification approach for IEM classification. Left: Mean ROC curves for each class. Right: Pairwise cosine distance between test embeddings used to evaluate the performance.

Standardizing an anatomical representation is not trivial due to the great anatomical variability present in our dataset which is two-fold. The first reason is that the LAA is one of the anatomical structures with more interpersonal variability, as it is present in many different shapes and sizes. The second reason is that most of the patients undergoing LAAO are elderly and usually suffer from various conditions that can affect the morphology and appearance of the heart.

We found a standardized representation of the complex anatomy that captures the challenges involved in the procedure. To do so, we focus on the analysis of how the anatomical centerline shown in Figure 4.9 that mimics the expected catheter trajectory can characterize the morphology surrounding the catheter path and find indicators of a higher risk of PDL.

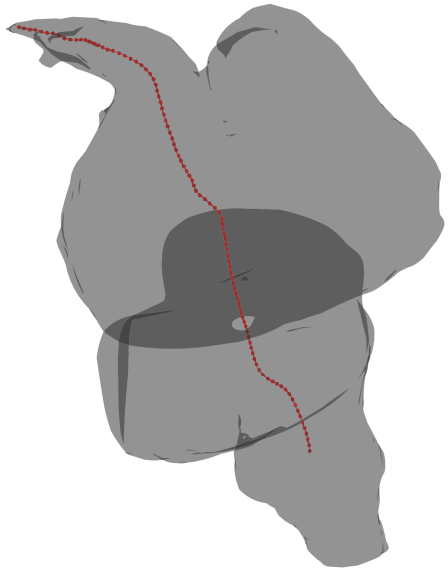


Figure 4.9: Representation of the extracted graph in red, in gray the unified 3D surface generated from the IVC, RA, LA, and LAA with the simulation of the FO opening.

In Contribution *G* we built morphology-aware graphs that follow the anatomical centerline of the expected catheter path. These graphs contain computed features that add information about the 3D

shape appearance in each cross-section along the centerline, in a way, we create a synthesized descriptor of a given anatomy. This allows us to build a structured graph that presents multiple technical benefits. We trained a GAT model for binary classification of PDL occurrence. We were not primarily interested in obtaining a high accuracy as, working with clinicians, we have observed that there is limited value in providing just a number that indicates the probability of a complication as the prognosis of a procedure. Therefore, we focused on analyzing the attention scores of the multi-head architecture once the model has learned how to detect these cases with some confidence. It can be seen in Figure 4.10 that the attention scores point us towards the angle of the IVC entering the heart as the area of higher interest. This is interesting as this area is usually not studied or analyzed given that it is further away from the key anatomical structure of the LAA.

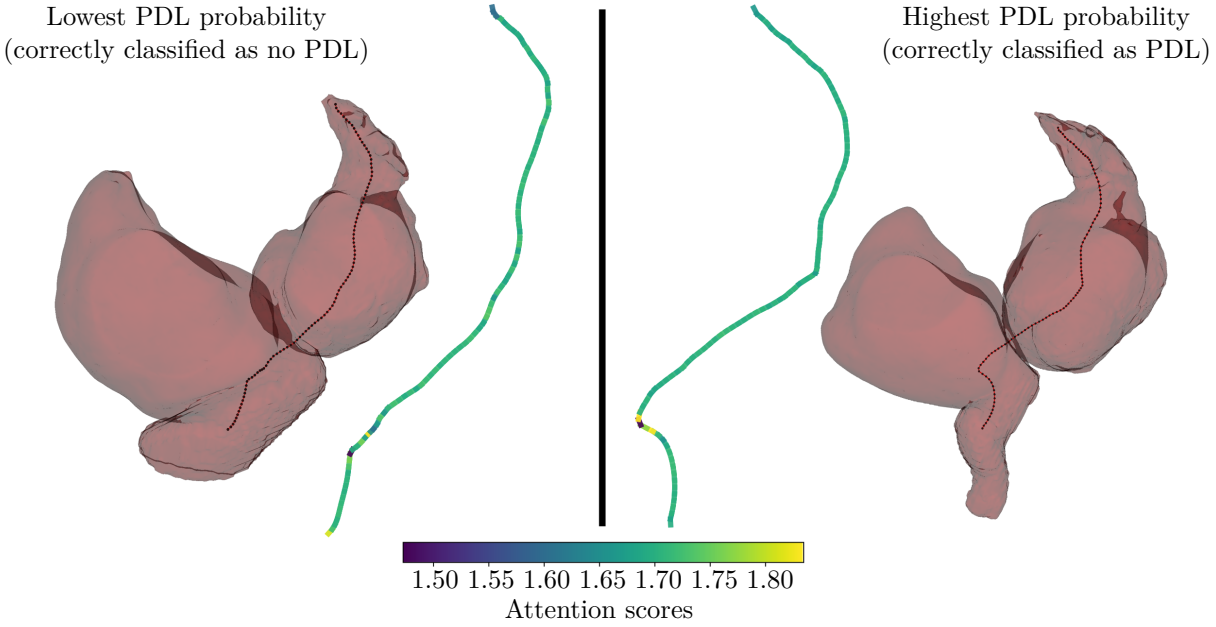


Figure 4.10: Attention scores of the cases from the test set with the lowest (left) and highest (right) estimated probability of PDL from our GAT model. The region of the constructed graphs with the highest change in attention scores is in the region between the IVC, RA, and FO, which is highly influenced by the orientation of the ICV.

This reveals the importance of morphological descriptors that successfully capture key elements and are well-designed. It is interesting how explainable approaches can also be used to open new clinical discussions and generate new research avenues and different hypotheses. This allows us to observe underlying trends that would otherwise be very difficult to analyze given the high complexity of the studied anatomy and the significant inter-personal morphological variability. All in all, we were able to use a GAT model trained on morphology-descriptive graphs to predict PDLs, using the attention scores to highlight the regions of our graph that have a greater impact on the algorithm's decision.

CHAPTER 5

Discussion

In this work, we have presented multiple novel methods applied to 3D image analysis for the clinical practice of medical implantable devices. Our research has been characterized by addressing complex, new, and interesting clinical questions with technically novel and powerful DL methods.

In the context of cochlear implant research, the application of image analysis methods to this anatomy has historically been quite limited compared to their extensive use in the study and treatment of other pathologies or anatomical structures. This gap is particularly important given the complexity and relevance of the inner ear's anatomy in the success of cochlear implantation therapy. Our research marks a significant advancement in this field, as we have developed the first automatic image-based approach specifically designed for the detection and further classification of IEMs. This innovation represents a critical step forward in the customization of hearing therapies, ultimately improving patient outcomes for this specific and challenging group. Besides the technical achievements of this thesis, by tackling clinical challenges on the inner ear anatomy, we have contributed to raising awareness within the medical imaging community about the critical importance of this field. We believe there is still a long way to support clinicians and gain deeper insights into cochlear implant therapy that can lead to more personalized and effective treatment options for individuals with hearing loss.

The first part of our research has centered on the use of anatomical landmarks. One of the main advantages of relying on landmarks is their faster annotation process making them less costly to obtain compared to other medical imaging annotations. Landmark-based approaches align well with current clinical practices in surgical planning, where specific landmarks and the distances between them are usually calculated to guide interventions. Clinicians' familiarity with anatomical landmarks and their design makes the inter-disciplinary collaboration smoother. When using landmarks, we can effectively parameterize specific aspects of an anatomical structure using a limited set of values. However, the utility of this parameterization is heavily dependent on the quality of the chosen landmarks and the specific clinical task.

In this project, we have used a DRL approach to localize the landmarks in 3D CT images. The human-like decision-making process of DRL models allows for more adaptive and robust handling of heterogeneous and challenging datasets as the ones utilized in this thesis. We proved the usability of this landmark-based approach for challenging tasks such as in the identification of nerves in clinical CT scans which is shown in Contribution *A*. Another advantage of landmark-based methods is their potential for producing direct anatomical measurements. However, this approach highly relies on the distinctiveness of the selected landmarks and the accuracy of the localization method. If the landmarks present variability, the resulting measurements can become unstable. To mitigate this, we often utilize landmarks as an initialization step for more advanced methods, as shown in Contributions *A* and *B*. In Contribution *A* we also showed that classical machine learning methods can still be used to characterize challenging structures with a landmark-based initialization using DRL. Furthermore, we have also used the landmark distribution for identifying specific subgroups, more specifically out-of-distribution samples, as explored in Contributions *C*, *D*, and *E*.

In these latter contributions, we further investigate a landmark-based approach for anomaly detection, utilizing both the implicit and explicit information encoded within the DRL landmark localization models. By parameterizing this approach for anomaly detection, we achieve superior performance compared to other self-supervised methods, which tend to be more sensitive to variations in image appearance. Nevertheless, our approach also requires an annotated dataset of a healthy population including anatomical landmarks which can be a limiting factor. This research shows the versatility and

effectiveness of landmark-based methods for anomaly detection in the medical imaging field. The need for robust and reliable anomaly detection algorithms in clinical practice cannot be overstated. Accurate detection of anomalies in clinical images is crucial for early diagnosis and personalized surgical treatment, making the development of such algorithms a top priority in the field. In our research, we have demonstrated that it is possible to extract valuable anomaly metrics from models originally trained for other tasks, such as landmark localization. This approach is particularly interesting because it uses an existing deployed model beyond its initial, task-specific, purpose.

Modeling the normal distribution of clinical data can be challenging, particularly when data availability is limited and the data itself is heterogeneous, which is the case for the data employed in this thesis and it is also a very common scenario in the medical image analysis field. Accessing multi-domain data from various acquisition centers improves the statistical power and representation of specific patient groups helping build more representative models but it also constitutes an additional challenge. Using this type of data can lead to more accurate models and better generalization across different populations. However, the sharing of medical images and clinical information remains a significant challenge due to strict privacy protection regulations and the lack of inter-center collaborations. Recent advances in federated learning offer promising solutions to these challenges by enabling collaborative model training across multiple sites without the need for direct data sharing. Federated learning techniques can handle domain adaptation issues, such as the variability in CT image appearance across different centers. Despite these advances, federated learning still relies at some level on cross-site collaboration. This requirement proves the ongoing need for cooperative efforts in the medical community to build models that can fully represent the data distribution and subgroups of patients in real clinical practice. While there are still obstacles to overcome, especially regarding data privacy and inter-center collaboration, the integration of anomaly detection into existing models and the adoption of federated learning approaches represent significant steps forward in improving clinical imaging and patient care.

Working with minoritarian pathologies presents significant challenges, particularly when the pathological spectrum is broad and highly imbalanced, as in our case. In such scenarios, traditional performance metrics, like high accuracy, can be misleading and do not necessarily reflect the true effectiveness of a model. This is because accuracy alone might mask poor performance on less frequent classes within the dataset. Therefore, it's crucial to consider performance metrics that more accurately reflect the model's ability to generalize across the full spectrum of the pathology and work towards explainable models that can indicate some of the reasoning behind the models' output.

We argue that supervised learning methods may not be the most suitable approach for handling this type of complex and imbalanced data, even with loss functions or training methods that are tailored for unbalanced data. In these scenarios, shortcut learning, where models rely on fictitious correlations rather than genuinely understanding the data, is a common issue that needs to be addressed. This phenomenon can lead to models that perform well on certain parts of the data but fail to generalize effectively across the broader spectrum of cases.

As demonstrated in Contribution *F*, we advocate for the development of robust structured latent spaces that can represent the variability of the data in a more organized manner while limiting the influence of the overall appearance of the input. To address the challenges posed by our dataset, we opted to use a point cloud representation of the various cochlear morphologies as input for our unsupervised classifier, rather than relying on the entire 3D image. This approach helped to avoid feeding the model with images from vastly different sources, which could have introduced unnecessary variability and affected the model's performance. Building these robust structured latent spaces in an unsupervised or self-supervised manner offers several advantages. First, it facilitates further classification by enabling clustering within the multidimensional space as shown in Contribution *F*, allowing for a more natural categorization of the data. It is particularly interesting to observe how this approach applied to the congenital malformations spectrum showed the continuity of the pathology and not only clearly separated clusters allowing to individually characterize every unique patient according to its morphological features. Second, these latent spaces can serve as foundational models for a range of other tasks related to the morphology of the anatomy, providing a flexible and adaptable framework for future research and applications. This approach addresses the challenges of working with minoritarian and heterogeneous pathologies as well as paves the way for more reliable and generalizable models in the field of medical

image analysis.

In our research, we have addressed both clinical and technically challenging questions, striving not only to develop methods that replicate clinicians' diagnostic processes but also to gain a deeper understanding of the underlying problems and assess the validity of clinical hypotheses. This approach is particularly evident in our final contribution (Paper *G*), where we aimed to evaluate whether the morphology of a patient's anatomy could be linked to the likelihood of a known post-procedural complication. Nowadays, there is still a tendency in the deep learning field to input all available data into models in pursuit of high-performance metrics. However, our work demonstrates the value of simplifying the input to focus only on the essential features that accurately capture the relevant variability within the dataset. This simplification allows for better understanding and control over our models, as it reduces complexity and potential noise that could obscure meaningful patterns, especially with small datasets. This can be seen in Contribution *G*, where we constructed simplified graphs that represent the anatomy involved in the catheter trajectory within the heart structure. These graphs, while simplified, effectively capture key aspects of the procedure's complexity. By using these standardized graphs across the population, we are able to analyze attention scores and enhance the explainability of our models. On the other hand, using full 3D images of the heart and surrounding structures would introduce additional complexity and variability, making it harder to draw clear conclusions.

Our close collaboration with clinicians has been a fundamental part of this work, as we have been actively involved in their daily practice and focused on addressing clinically relevant research questions. This collaboration led to the novel finding that the angle of the IVC may be more closely related to post-procedural leaks than previously thought. This discovery is particularly interesting because this anatomical feature is often overlooked in pre-procedural analysis and planning. The implications of this finding suggest that it may be worthwhile to begin incorporating the analysis of the IVC into the pre-procedural evaluation analysis, as its morphology could potentially limit the trajectories available during the procedure and thus influence patient outcomes. This shows the importance of our approach in, not only developing technically robust models but also, contributing with new insights that can directly impact clinical practice.

Regarding future research directions in the context of cochlear implants, there is a clear need for surgical planning support, particularly for cases involving congenital malformations. We believe our work represents the initial steps of a longer journey. Analyzing the variability in anatomy and understanding how it differs from normal anatomical cases will be crucial for future developments. Identifying surgically relevant points of interest that constrain the insertion trajectory is also essential. Given the complexity of these cases, close collaboration between technical and clinical experts will be required. Additionally, analyzing postoperative CT images of patients with IEMs could provide valuable insights into the type of insertion and potential surgical trajectories.

In the context of the LAA and the risk of PDL, further analysis of the IVC's angle should be conducted to fully understand how this anatomical feature influences the procedure. Moreover, post-procedural analysis of the final device placement could determine whether the IVC's angle is associated with the coaxiality of the device within the LAA. There are also numerous other factors that influence PDL occurrence, as previously mentioned. It would be highly beneficial to incorporate additional contributing factors derived from preoperative CT images into our approach, potentially including demographic data such as gender, age, or systolic blood pressure, to obtain a better understanding of the clinical case.

CHAPTER 6

Conclusions

In this thesis, the primary contributions generated during the 3-year PhD program have been presented and thoroughly discussed. These contributions are the result of a collaborative, multi-disciplinary effort between various research teams, reflecting the intersection of technical innovation and clinical application in a medical imaging context.

One of the main research venues has been the development and application of DRL for landmark detection within the inner ear anatomy. Papers *A* and *B* are novel examples of how DRL can be leveraged to identify and utilize reliable anatomical landmarks, enabling the characterization of challenging structures and the precise identification of regions of interest. Specifically, these methods have been used to characterize neural structures and determine the patient's head orientation in CT scans with great success.

Furthermore, this thesis introduces the first method for detecting congenital malformations in the inner ear, as presented in Contributions *C*, *D*, and *E*. This approach has demonstrated superior performance compared to existing state-of-the-art methods for anomaly detection. Building on top of the DRL model introduced in Paper *A*, the work derives both implicit and explicit measurements from the trained model that can be used for reliable anomaly detection, utilizing not only the output distribution but also the intrinsic properties of the DRL model itself.

The first automated approach for classifying congenital inner ear anomalies is introduced in this dissertation, in Contribution *F*. It is an unsupervised approach that is used for latent space representation of anatomically describing point clouds. This novel approach lays the groundwork for supporting clinicians in the diagnosis of complex and rare inner ear pathologies, offering a method that could be adapted to other conditions where structural abnormalities are a key diagnostic factor.

Finally, in Contribution *G* we developed an innovative approach to characterize the complexity of LAAO procedures. This was achieved by generating a graph that captures the catheter trajectory's centerline and describes the surrounding anatomy, to evaluate its influence on the occurrence of PDLs. By training a GAT for classification and analyzing the attention score distribution within the graph, we show new interesting results that could lead to further research and potential improvements in procedural outcomes. In this work, we show the potential of new explainability methods to contribute to clinical research.

This thesis demonstrates the importance of designing robust and technically advanced methods to address specific clinical challenges, particularly those related to medical device insertion procedures. These interventions, which demand a high level of skill and expertise from clinicians, also require domain-specific tools designed to support clinical practice. Our research has focused on improving procedural planning and prognosis in this area, with the ultimate aim of facilitating safer and more effective medical device insertion procedures.

Bibliography

[1] Y. Akita, R. Nakayama, A. Hizukuri, and S. Kido. "Anomaly Detection for Lung CT image Using Efficient GAN with Learning Stabilizations." In: *2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCISISIS)*. 2022, pages 1–2. DOI: 10.1109/SCISISIS55246.2022.10001896 (cited on page 16).

[2] A. Alansary, O. Oktay, Y. Li, L. L. Folgoc, B. Hou, G. Vaillant, K. Kamnitsas, A. Vlontzos, B. Glocker, B. Kainz, and D. Rueckert. "Evaluating reinforcement learning agents for anatomical landmark detection." In: *Medical Image Analysis* 53 (April 2019), pages 156–164. ISSN: 13618423. DOI: 10.1016/j.media.2019.02.007 (cited on page 14).

[3] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. "A brief survey of deep reinforcement learning." In: *arXiv preprint arXiv:1708.05866* (2017) (cited on page 12).

[4] M. Astaraki, Ö. Smedby, and C. Wang. "Prior-aware autoencoders for lung pathology segmentation." In: *Medical Image Analysis* 80 (2022), page 102491 (cited on page 15).

[5] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni. "Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study." In: *Medical Image Analysis* 69 (2021), page 101952. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2020.101952. URL: https://www.sciencedirect.com/science/article/pii/S1361841520303169 (cited on page 15).

[6] C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab. "Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2020, pages 718–727 (cited on page 16).

[7] C. I. Bercea, B. Wiestler, D. Rueckert, and J. A. Schnabel. "Towards Universal Unsupervised Anomaly Detection in Medical Imaging." In: *arXiv preprint arXiv:2401.10637* (2024) (cited on page 16).

[8] S. Berrettini, A. de Vito, L. Bruschini, S. Passetti, and F. Forli. "Facial nerve stimulation after cochlear implantation: Our experience; [La stimolazione del nervo facciale dopo la procedura di impianto cocleare: La nostra esperienza]." In: *Acta Otorhinolaryngologica Italica* 31.1 (2011). Cited by: 47, pages 11–16. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-79957791223&partnerID=40&md5=cad551937c2838ce45b9b538e74a8571 (cited on page 5).

[9] D. C. Bigelow, D. J. Kay, K. O. Rafter, M. Montes, G. W. Knox, and D. M. Yousem. "Facial nerve stimulation from cochlear implants." In: *American Journal of Otology* 19.2 (1998), pages 163–169 (cited on page 5).

[10] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, and P. Cui. "Structural Deep Clustering Network." In: *Proceedings of The Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, 2020, pages 1400–1410. ISBN: 9781450370233. DOI: 10.1145/3366423.3380214. URL: https://doi.org/10.1145/3366423.3380214 (cited on page 18).

[11] S. Brody, U. Alon, and E. Yahav. "How attentive are graph attention networks?" In: *arXiv preprint arXiv:2105.14491* (2021) (cited on page 18).

[12] D. Brotto, F. Sorrentino, R. Cenedese, I. Avato, R. Bovo, P. Trevisi, and R. Manara. "Genetics of inner ear malformations: a review." In: *Audiology Research* 11.4 (2021), pages 524–536 (cited on page 6).

[13] L. Cao, S. Asadi, W. Zhu, C. Schmidli, and M. Sjöberg. "Simple, Scalable, and Stable Variational Deep Clustering." In: *Machine Learning and Knowledge Discovery in Databases*. Edited by F. Hutter, K. Kersting, J. Lijffijt, and I. Valera. Cham: Springer International Publishing, 2021, pages 108–124. ISBN: 978-3-030-67658-2 (cited on page 18).

[14] T. Chan and L. Vese. "An active contour model without edges." In: *International conference on scale-space theories in computer vision*. Springer. 1999, pages 141–151 (cited on page 29).

[15] X. Chen, S. You, K. C. Tezcan, and E. Konukoglu. "Unsupervised lesion detection via image restoration with a normative prior." In: *Medical image analysis* 64 (2020), page 101713 (cited on page 15).

[16] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier. "Development and validation of deep learning algorithms for detection of critical findings in head CT scans." In: *arXiv preprint arXiv:1803.05854* (2018) (cited on page 24).

[17] S. Dazert, J. P. Thomas, A. Loth, T. Zahnert, and T. Stöver. "Cochlear Implantation: Diagnosis, Indications, and Auditory Rehabilitation Results." In: *Dtsch Arztebl International* 117.41 (2020), pages 690–700. DOI: 10.3238/arztebl.2020.0690. eprint: https://www.aerzteblatt.de/pdf.asp?id=216085 (cited on page 4).

[18] B. De Foer, C. Kenis, D. Van Melkebeke, J.-P. Vercruysse, T. Somers, M. Pouillon, E. Offeciers, and J. W. Casselman. "Pathology of the vestibulocochlear nerve." In: *European Journal of Radiology* 74.2 (2010). Imaging of Cranial Nerves and Brachial Plexus, pages 349–358. ISSN: 0720-048X. DOI: https://doi.org/10.1016/j.ejrad.2009.06.033. URL: https://www.sciencedirect.com/science/article/pii/S0720048X10000677 (cited on page 5).

[19] L. De Raeve. "Cochlear implants in Belgium: Prevalence in paediatric and adult cochlear implantation." In: *European Annals of Otorhinolaryngology, Head and Neck Diseases* 133 (2016). 12th European Symposium on Pediatric Cochlear Implant (ESPCI 2015), S57–S60. ISSN: 1879-7296. DOI: https://doi.org/10.1016/j.anorl.2016.04.018. URL: https://www.sciencedirect.com/science/article/pii/S1879729616300813 (cited on page 5).

[20] A. E. Dhanasingh, N. M. Weiss, V. Erhard, F. Altamimi, P. Roland, A. Hagr, V. Van Rompaey, and P. Van de Heyning. "A novel three-step process for the identification of inner ear malformation types." In: *Laryngoscope Investigative Otolaryngology* 7.6 (2022), pages 2020–2028 (cited on page 7).

[21] P. L. Diez, K. A. Juhl, J. V. Sundgaard, H. Diab, J. Margeta, F. Patou, and R. R. Paulsen. "Deep reinforcement learning for detection of abnormal anatomies." In: *Proceedings of the Northern Lights Deep Learning Workshop*. Volume 3. 2022 (cited on page v).

[22] P. L. Diez, J. V. Sundgaard, J. Margeta, K. Diab, F. Patou, and R. R. Paulsen. "Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images." In: *Computerized Medical Imaging and Graphics* 113 (2024), page 102343 (cited on page v).

[23] E. W. Dijkstra. "A note on two problems in connexion with graphs." In: *Numerische mathematik* 1.1 (1959), pages 269–271 (cited on page 29).

[24] S. F. Frisken. "Surfacenets for multi-label segmentations with preservation of sharp boundaries." In: *The Journal of computer graphics techniques* 11.1 (2022), page 34 (cited on page 17).

[25] T. Furuya and R. Ohbuchi. "DeepDiffusion: Unsupervised Learning of Retrieval-Adapted Representations via Diffusion-Based Ranking on Latent Feature Manifold." In: *IEEE Access* 10 (2022), pages 116287–116301. DOI: 10.1109/ACCESS.2022.3218909 (cited on page 34).

[26] B. M. Gare, T. Hudson, S. A. Rohani, D. G. Allen, S. K. Agrawal, and H. M. Ladak. "Multi-atlas segmentation of the facial nerve from clinical CT for virtual reality simulators." In: *International Journal of Computer Assisted Radiology and Surgery* 15 (2 February 2020), pages 259–267. ISSN: 18616429. DOI: 10.1007/s11548-019-02091-0 (cited on page 13).

[27] L. Gärtner, B. C. Backus, N. Le Goff, A. Morgenstern, T. Lenarz, and A. Büchner. "Cochlear Implant Stimulation Parameters Play a Key Role in Reducing Facial Nerve Stimulation." In: *Journal of Clinical Medicine* 12.19 (2023), page 6194 (cited on page 5).

[28] F. C. Ghesu, B. Georgescu, S. Grbic, A. K. Maier, J. Hornegger, and D. Comaniciu. "Robust Multi-scale Anatomical Landmark Detection in Incomplete 3D-CT Data." In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*. Edited by M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne. Cham: Springer International Publishing, 2017, pages 194–202. ISBN: 978-3-319-66182-7 (cited on page 14).

[29] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu. "An Artificial Agent for Anatomical Landmark Detection in Medical Images." In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*. Edited by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells. Cham: Springer International Publishing, 2016, pages 229–237. ISBN: 978-3-319-46726-9 (cited on page 14).

[30] F. C. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu. "Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (1 January 2019), pages 176–189. ISSN: 19393539. DOI: 10.1109/TPAMI.2017.2782687 (cited on page 14).

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets." In: *Advances in neural information processing systems* 27 (2014) (cited on page 15).

[32] M. Grewal, T. M. Deist, J. Wiersma, P. A. N. Bosman, and T. Alderliesten. "An end-to-end deep learning approach for landmark detection and matching in medical images." In: SPIE-Intl Soc Optical Eng, March 2020, page 79. DOI: 10.1117/12.2549302 (cited on page 13).

[33] J. J. Han, M.-W. Suh, M. K. Park, J.-W. Koo, J. H. Lee, and S. H. Oh. "A predictive model for cochlear implant outcome in children with cochlear nerve deficiency." In: *Scientific reports* 9.1 (2019), page 1154 (cited on page 5).

[34] F. Heutink, V. Koch, B. Verbist, W. J. van der Woude, E. Mylanus, W. Huinck, I. Sechopoulos, and M. Caballo. "Multi-Scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution CT images." In: *Computer methods and programs in biomedicine* 191 (2020), page 105387 (cited on page 14).

[35] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang. "Graphlime: Local interpretable model explanations for graph neural networks." In: *IEEE Transactions on Knowledge and Data Engineering* 35.7 (2022), pages 6968–6972 (cited on page 19).

[36] K. Kaga. *Cochlear implantation in children with inner ear malformation and cochlear nerve deficiency*. Springer, 2016 (cited on page 7).

[37] D. C. Kelsall, J. K. Shallop, T. G. Brammeier, and E. C. Prenger. "Facial nerve stimulation after Nucleus 22-channel cochlear implantation." In: *Otology & Neurotology* 18.3 (1997), pages 336–341 (cited on page 5).

[38] D. Kim and A. Oh. "How to find your friendly neighborhood: Graph attention design with self-supervision." In: *arXiv preprint arXiv:2204.04879* (2022) (cited on page 18).

[39] D. Kingma. "Auto-Encoding Variational Bayes." In: *arXiv preprint arXiv:1312.6114* (2013) (cited on page 15).

[40] D. Lakkireddy, J. E. Nielsen-Kudsk, S. Windecker, D. Thaler, M. J. Price, A. Gambhir, N. Gupta, K. Koulogiannis, L. Marcoff, A. Mediratta, J. A. Anderson, R. Gage, and C. R. Ellis. "Mechanisms, predictors, and evolution of severe peri-device leaks with two different left atrial appendage occluders." In: *EP Europace* 25.9 (August 2023), euad237. ISSN: 1099-5129. DOI: 10.1093/europace/euad237. eprint: https://academic.oup.com/europace/article-pdf/25/9/euad237/51114353/euad237.pdf. URL: https://doi.org/10.1093/europace/euad237 (cited on page 9).

[41] J. Lee, I. Lee, and J. Kang. "Self-attention graph pooling." In: *International conference on machine learning*. pmlr. 2019, pages 3734–3743 (cited on page 19).

[42] S. M. Lee, H. P. Kim, K. Jeon, S. H. Lee, and J. K. Seo. "Automatic 3D cephalometric annotation system using shadowed 2D image-based machine learning." In: *Physics in Medicine and Biology* 64 (5 2019). ISSN: 13616560. DOI: 10.1088/1361-6560/ab00c9 (cited on page 14).

[43] G. Leroy, D. Rueckert, and A. Alansary. "Communicative Reinforcement Learning Agents for Landmark Detection in Brain Images." In: *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*. Edited by S. M. Kia, H. Mohy-ud-Din, A. Abdulkadir, C. Bass, M. Habes, J. M. Rondina, C. Tax, H. Wang, T. Wolfers, S. Rathore, and M. Ingalhalikar. Cham: Springer International Publishing, 2020, pages 177–186. ISBN: 978-3-030-66843-3 (cited on pages 13, 28, 29).

[44] Y. Li, A. Alansary, J. J. Cerrolaza, B. Khanal, M. Sinclair, J. Matthew, C. Gupta, C. Knight, B. Kainz, and D. Rueckert. "Fast multiple landmark localisation using a patch-based iterative network." In: volume 11070 LNCS. 2018. DOI: 10.1007/978-3-030-00928-1_64 (cited on page 14).

[45] P. López Diez, J. Margeta, K. Diab, F. Patou, and R. R. Paulsen. "Unsupervised classification of congenital inner ear malformations using DeepDiffusion for latent space representation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pages 652–662 (cited on page v).

[46] P. López Diez, K. Sørensen, J. V. Sundgaard, K. Diab, J. Margeta, F. Patou, and R. R. Paulsen. "Deep reinforcement learning for detection of inner ear abnormal anatomy in computed tomography." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pages 697–706 (cited on page v).

[47] P. López Diez, J. V. Sundgaard, F. Patou, J. Margeta, and R. R. Paulsen. "Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer. 2021, pages 519–528 (cited on page v).

[48] W. E. Lorensen and H. E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm." In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '87. New York, NY, USA: Association for Computing Machinery, 1987, pages 163–169. ISBN: 0897912276. DOI: 10.1145/37401.37422. URL: https://doi.org/10.1145/37401.37422 (cited on page 17).

[49] D. Ludhwani and J. S. Wieters. "Paroxysmal atrial fibrillation." In: (2018) (cited on page 8).

[50] S. M. Lundberg and S.-I. Lee. "A unified approach to interpreting model predictions." In: *Advances in neural information processing systems* 30 (2017) (cited on page 19).

[51] Q. Ma, E. Kobayashi, B. Fan, K. Nakagawa, I. Sakuma, K. Masamune, and H. Suenaga. "Automatic 3D landmarking model using patch-based deep neural networks for CT image of oral and maxillofacial surgery." In: *International Journal of Medical Robotics and Computer Assisted Surgery* 16 (3 June 2020). ISSN: 1478596X. DOI: 10.1002/rcs.2093 (cited on page 14).

[52] J. Margeta, R. Hussain, P. L. Diez, A. Morgenstern, T. Demarcy, Z. Wang, D. Gnansia, O. M. Manzanera, C. Vandersteen, H. Delingette, et al. "A Web-Based Automated Image Processing Research Platform for Cochlear Implantation-Related Studies." In: *Journal of Clinical Medicine* 11.22 (2022), page 6640 (cited on page vi).

[53] J. Margeta, R. Hussain, P. López Diez, A. Morgenstern, T. Demarcy, Z. Wang, D. Gnansia, O. Martinez Manzanera, C. Vandersteen, H. Delingette, A. Buechner, T. Lenarz, F. Patou, and N. Guevara. "A Web-Based Automated Image Processing Research Platform for Cochlear Implantation-Related Studies." In: *Journal of Clinical Medicine* 11.22 (2022). ISSN: 2077-0383. DOI: 10.3390/jcm11226640. URL: https://www.mdpi.com/2077-0383/11/22/6640 (cited on page 14).

[54] V. Mnih. "Playing atari with deep reinforcement learning." In: *arXiv preprint arXiv:1312.5602* (2013) (cited on page 12).

[55] M. d. A. Moraes, P. A. P. d. Jesus, L. S. Muniz, G. A. Costa, L. V. Pereira, L. M. Nascimento, C. A. d. S. Teles, C. A. Baccin, and F. C. Mussi. "Ischemic stroke mortality and time for hospital arrival: analysis of the first 90 days." In: *Revista da Escola de Enfermagem da USP* 57 (2023), e20220309 (cited on page 8).

[56] N. Mrabah, M. Bouguessa, and R. Ksantini. "Adversarial Deep Embedded Clustering: On a Better Trade-off Between Feature Randomness and Feature Drift." In: *IEEE Transactions on Knowledge and Data Engineering* 34.4 (2022), pages 1603–1617. DOI: 10.1109/TKDE.2020.2997772 (cited on page 18).

[57] J. M. H. Noothout, B. D. de Vos, J. M. Wolterink, T. Leiner, and I. Isgum. "CNN-based Landmark Detection in Cardiac CTA Scans." In: *CoRR* abs/1804.04963 (2018). arXiv: 1804.04963. URL: http://arxiv.org/abs/1804.04963 (cited on pages 13, 14).

[58] N. Pawlowski, M. C. Lee, M. Rajchl, S. McDonagh, E. Ferrante, K. Kamnitsas, S. Cooke, S. Stevenson, A. Khetani, T. Newman, et al. "Unsupervised lesion detection in brain CT using bayesian convolutional autoencoders." In: (2018) (cited on page 15).

[59] C. Payer, D. Štern, H. Bischof, and M. Urschler. "Regressing heatmaps for multiple landmark localization using CNNs." In: volume 9901 LNCS. Springer Verlag, 2016, pages 230–238. ISBN: 9783319467221. DOI: 10.1007/978-3-319-46723-8_27 (cited on page 14).

[60] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi. "Deep subspace clustering with sparsity prior." In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA: AAAI Press, 2016, pages 1925–1931. ISBN: 9781577357704 (cited on page 18).

[61] W. H. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, et al. "Fast unsupervised brain anomaly detection and segmentation with diffusion models." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pages 705–714 (cited on page 16).

[62] W. H. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso. "Unsupervised brain imaging 3D anomaly detection and segmentation with transformers." In: *Medical Image Analysis* 79 (2022), page 102475 (cited on page 15).

[63] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. "Pointnet: Deep learning on point sets for 3d classification and segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 652–660 (cited on page 17).

[64] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." In: *Advances in neural information processing systems* 30 (2017) (cited on page 17).

[65] A.-T. Radutoiu, F. Patou, J. Margeta, R. R. Paulsen, and P. López Diez. "Accurate localization of inner ear regions of interests using deep reinforcement learning." In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2022, pages 416–424 (cited on page v).

[66] H. Rask-Andersen, W. Liu, E. Erixon, A. Kinnefors, K. Pfaller, A. Schrott-Fischer, and R. Glueckert. "Human Cochlea: Anatomical Characteristics and their Relevance for Cochlear Implantation." In: *The Anatomical Record* 295.11 (2012), pages 1791–1811. DOI: https://doi.org/10.1002/ar.22599. eprint: https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/ar.22599. URL: https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/ar.22599 (cited on page 20).

[67] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P. S. Yu, and L. He. "Deep Clustering: A Comprehensive Survey." In: *IEEE Transactions on Neural Networks and Learning Systems* (2024), pages 1–21. DOI: 10.1109/TNNLS.2024.3403155 (cited on page 17).

[68] J. Rk. "Congenital malformations of the inner ear: a classification based on embryogenesis." In: *Laryngoscope* 97.40 (1987), pages 2–14 (cited on pages 6, 7).

[69] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: (May 2015). URL: http://arxiv.org/abs/1505.04597 (cited on page 14).

[70] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pages 234–241 (cited on pages 14, 34).

[71] D. Sato, S. Hanaoka, Y. Nomura, T. Takenaga, S. Miki, T. Yoshikawa, N. Hayashi, and O. Abe. "A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes." In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Volume 10575. SPIE. 2018, pages 388–393 (cited on pages 15, 32).

[72] J. Saw, P. Fahmy, P. DeJong, M. Lempereur, R. Spencer, M. Tsang, K. Gin, J. Jue, J. Mayo, P. McLaughlin, and S. Nicolaou. "Cardiac CT angiography for device surveillance after endovascular left atrial appendage closure." In: *European Heart Journal - Cardiovascular Imaging* 16.11 (April 2015), pages 1198–1206. ISSN: 2047-2404. DOI: 10.1093/ehjci/jev067. URL: https://doi.org/10.1093/ehjci/jev067 (cited on page 26).

[73] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth. "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks." In: *Medical image analysis* 54 (2019), pages 30–44 (cited on page 16).

[74] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery." In: *International conference on information processing in medical imaging*. Springer. 2017, pages 146–157 (cited on page 16).

[75] W. Schroeder, R. Maynard, and B. Geveci. "Flying edges: A high-performance scalable isocontouring algorithm." In: *2015 IEEE 5th Symposium on Large Data Analysis and Visualization (LDAV)*. IEEE. 2015, pages 33–40 (cited on page 17).

[76] L. Sennaroğlu and M. D. Bajin. "Classification and current management of inner ear malformations." In: *Balkan medical journal* 34.5 (2017), pages 397–411 (cited on page 6).

[77] K. Sørensen, O. Camara, O. d. Backer, K. F. Kofoed, and R. R. Paulsen. "NUDF: Neural Unsigned Distance Fields for High Resolution 3D Medical Image Segmentation." In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. 2022, pages 1–5. DOI: 10.1109/ISBI52829.2022.9761610 (cited on page 26).

[78] C. Sun, C. Li, X. Lin, T. Zheng, F. Meng, X. Rui, and Z. Wang. "Attention-based graph neural networks: a survey." In: *Artificial Intelligence Review* 56.Suppl 2 (2023), pages 2263–2310 (cited on page 19).

[79] L. Sun, J. Wang, Y. Huang, X. Ding, H. Greenspan, and J. Paisley. "An adversarial learning approach to medical image synthesis for lesion detection." In: *IEEE journal of biomedical and health informatics* 24.8 (2020), pages 2303–2314 (cited on page 16).

[80] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic attribution for deep networks." In: *International conference on machine learning*. PMLR. 2017, pages 3319–3328 (cited on page 19).

[81] F. Toulgoat, J. Sarrazin, F. Benoudiba, Y. Pereon, E. Auffray-Calvier, B. Daumas-Duport, A. Lintia-Gaultier, and H. Desal. "Facial nerve: From anatomy to pathology." In: *Diagnostic and Interventional Imaging* 94.10 (2013). Imaging of the cranial and peripheral nerves, pages 1033–1042. ISSN: 2211-5684. DOI: https://doi.org/10.1016/j.diii.2013.06.016. URL: https://www.sciencedirect.com/science/article/pii/S2211568413002155 (cited on page 4).

[82] M.-P. Tuset, A. Baptiste, F. Cyna Gorse, O. Sterkers, Y. Nguyen, G. Lahlou, E. Ferrary, and I. Mosnier. "Facial nerve stimulation in adult cochlear implant recipients with far advanced otosclerosis." In: *Laryngoscope Investigative Otolaryngology* 8.1 (2023), pages 220–229 (cited on page 4).

[83] A. Van Horn, C. Hayden, A. D. Mahairas, P. Leader, and M. L. Bush. "Factors influencing aberrant facial nerve stimulation following cochlear implantation: a systematic review and meta-analysis." In: *Otology & Neurotology* 41.8 (2020), pages 1050–1059 (cited on page 4).

[84] G. van Tulder and M. de Bruijne. "Unpaired, unsupervised domain adaptation assumes your domains are already similar." In: *Medical Image Analysis* 87 (2023), page 102825. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2023.102825. URL: https://www.sciencedirect.com/science/article/pii/S1361841523000865 (cited on page 18).

[85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems* 30 (2017) (cited on page 18).

[86] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al. "Graph attention networks." In: *stat* 1050.20 (2017), pages 10–48550 (cited on page 18).

[87] J. F. Viles-Gonzalez, S. Kar, P. Douglas, S. Dukkipati, T. Feldman, R. Horton, D. Holmes, and V. Y. Reddy. "The clinical impact of incomplete left atrial appendage closure with the Watchman Device in patients with atrial fibrillation: a PROTECT AF (Percutaneous Closure of the Left Atrial Appendage Versus Warfarin Therapy for Prevention of Stroke in Patients With Atrial Fibrillation) substudy." In: *Journal of the American College of Cardiology* 59.10 (2012), pages 923–929 (cited on page 9).

[88] A. Vlontzos, A. Alansary, K. Kamnitsas, D. Rueckert, and B. Kainz. "Multiple landmark detection using multi-agent reinforcement learning." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pages 262–270 (cited on page 13).

[89] E. H. Voormolen, M. van Stralen, P. A. Woerdeman, J. P. Pluim, H. J. Noordmans, M. A. Viergever, L. Regli, and J. W. B. van der Sprenkel. "Determination of a facial nerve safety zone for navigated temporal bone surgery." In: *Neurosurgery* 70 (1 Suppl Operative 2012). ISSN: 15244040. DOI: 10.1227/NEU.0b013e31822e7fc3 (cited on page 13).

[90] S. Wachter, B. Mittelstadt, and C. Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." In: *Harv. JL & Tech.* 31 (2017), page 841 (cited on page 19).

[91] G. Wang, R. Ying, J. Huang, and J. Leskovec. "Improving graph attention networks with large margin-based constraints." In: *arXiv preprint arXiv:1910.11945* (2019) (cited on page 18).

[92] J. Wang and J. Jiang. "Unsupervised deep clustering via adaptive GMM modeling and optimization." In: *Neurocomputing* 433 (2021), pages 199–211. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.12.082. URL: https://www.sciencedirect.com/science/article/pii/S0925231220319949 (cited on page 18).

[93] Y. Wang, T. Lenarz, A. Kral, and S. John. "Automatic Landmark Localization in 3D Medical CT Images: Few-Shot Learning through Optimized Data Pre-Processing and Network Design." In: *Proceedings of the 2023 10th International Conference on Bioinformatics Research and Applications*. 2023, pages 1–7 (cited on page 14).

[94] Z. Wang, C. Vandersteen, C. Raffaelli, N. Guevara, F. Patou, and H. Delingette. "One-shot learning for landmarks detection." In: *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*. Springer. 2021, pages 163–172 (cited on page 14).

[95] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al. "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images." In: *Radiology: Artificial Intelligence* 5.5 (2023) (cited on page 26).

[96] G. Widmann, D. Dejaco, A. Luger, and J. Schmutzhard. "Pre-and post-operative imaging of cochlear implants: a pictorial review." In: *Insights into Imaging* 11.1 (2020), page 93 (cited on page 5).

[97] J. Wu, K. Long, F. Wang, C. Qian, C. Li, Z. Lin, and H. Zha. "Deep Comprehensive Correlation Mining for Image Clustering." In: *International Conference on Computer Vision*. 2019 (cited on page 18).

[98] N. C. Wunderlich, R. Beigel, M. J. Swaans, S. Y. Ho, and R. J. Siegel. "Percutaneous Interventions for Left Atrial Appendage Exclusion: Options, Assessment, and Imaging Using 2D and 3D Echocardiography." In: *JACC: Cardiovascular Imaging* 8.4 (2015), pages 472–488. ISSN: 1936-878X. DOI: https://doi.org/10.1016/j.jcmg.2015.02.002. URL: https://www.sciencedirect.com/science/article/pii/S1936878X15000935 (cited on page 8).

[99] Z. Xu, Q. Huang, J. Park, M. Chen, D. Xu, D. Yang, D. Liu, and S. K. Zhou. "Supervised Action Classifier: Approaching Landmark Detection as Image Partitioning." In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*. Edited by M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne. Cham: Springer International Publishing, 2017, pages 338–346. ISBN: 978-3-319-66179-7 (cited on page 14).

[100] J. Yang, D. Parikh, and D. Batra. "Joint Unsupervised Learning of Deep Representations and Image Clusters." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cited on page 18).

[101] L. Yang, W. Fan, and N. Bouguila. "Deep Clustering Analysis via Dual Variational Autoencoder With Spherical Latent Embeddings." In: *IEEE Transactions on Neural Networks and Learning Systems* 34.9 (2023), pages 6303–6312. DOI: 10.1109/TNNLS.2021.3135460 (cited on page 18).

[102] X. Yang, J. Yan, Y. Cheng, and Y. Zhang. "Learning Deep Generative Clustering via Mutual Information Maximization." In: *IEEE Transactions on Neural Networks and Learning Systems* 34 (2022), pages 6263–6275. URL: https://api.semanticscholar.org/CorpusID:245703352 (cited on page 18).

[103] X. Yang, C. Deng, F. Zheng, J. Yan, and W. Liu. "Deep Spectral Clustering Using Dual Autoencoder Network." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pages 4066–4075. DOI: 10.1109/CVPR.2019.00419. URL: http://openaccess.thecvf.com/content%5CCVPR%5C2019/html/Yang%5CDeep%5CSpectral%5CClustering%5CUsing%5CDual%5CAutoencoder%5CNetwork%5CCVPR%5C2019%5Cpaper.html (cited on page 18).

[104] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. "Gnnexplainer: Generating explanations for graph neural networks." In: *Advances in neural information processing systems* 32 (2019) (cited on page 19).

[105] H. Yuan, H. Yu, S. Gui, and S. Ji. "Explainability in graph neural networks: A taxonomic survey." In: *IEEE transactions on pattern analysis and machine intelligence* 45.5 (2022), pages 5782–5799 (cited on page 19).

[106] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability." In: *Neuroimage* 31.3 (2006), pages 1116–1128 (cited on page 22).

[107] D. Zhang, Y. Liu, J. H. Noble, and B. M. Dawant. "Automatic localization of landmark sets in head CT images with regression forests for image registration initialization." In: volume 9784. SPIE, March 2016, page 97841M. ISBN: 9781510600195. DOI: 10.1117/12.2216925 (cited on page 14).

[108] J. Zhang, M. Liu, and D. Shen. "Detecting Anatomical Landmarks from Limited Medical Imaging Data Using Two-Stage Task-Oriented Deep Neural Networks." In: *IEEE Transactions on Image Processing* 26 (10 2017). ISSN: 10577149. DOI: 10.1109/TIP.2017.2721106 (cited on page 14).

[109] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu. "3D deep learning for efficient and robust landmark detection in volumetric data." In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9349 (2015), pages 565–572. ISSN: 16113349. DOI: 10.1007/978-3-319-24553-9_69 (cited on page 14).

[110] D. Zhu, S. Chen, X. Ma, and R. Du. "Adaptive Graph Convolution Using Heat Kernel for Attributed Graph Clustering." In: *Applied Sciences* 10.4 (2020). ISSN: 2076-3417. DOI: 10.3390/app10041473. URL: https://www.mdpi.com/2076-3417/10/4/1473 (cited on page 18).

[111] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein. "Unsupervised anomaly localization using variational auto-encoders." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pages 289–297 (cited on page 15).

PAPER A

Facial and Cochlear Nerves Characterization Using Deep Reinforcement Learning for Landmark Detection

Authors Paula López Diez, Josefine Vilsbøll Sundgaard, François Patou, Jan Margeta, and Rasmus R. Paulsen.

Journal Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24

Year 2021

Status Published

DOI https://doi.org/10.1007/978-3-030-87202-1_50

Facial and Cochlear Nerves Characterization Using Deep Reinforcement Learning for Landmark Detection

Paula López Diez[1,2(✉)], Josefine Vilsbøll Sundgaard[1], François Patou[2], Jan Margeta[2,3], and Rasmus Reinhold Paulsen[1]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
plodi@dtu.dk
[2] Oticon Medical, Research and Technology Group, Smørum, Denmark
[3] KardioMe, Research and Development, Nova Dubnica, Slovakia

Abstract. We propose a pipeline for the characterization of facial and cochlear nerves in CT scans, a task specifically relevant for cochlear implant surgery planning. These structures are hard to locate in clinical CT scans due to their small size relative to the image resolution, the lack of contrast, and the proximity to other similar structures in this region. We define key landmarks around the facial and cochlear nerves and locate them using deep reinforcement learning with communicative multi-agents based on the C-MARL model. These landmarks are used as initialization for customized characterization methods. These include the automated direct measurement of the diameter of the cochlear nerve canal and extraction of the cochlear nerve cross-section followed by its segmentation using active contours. We also derive a path selection algorithm for optimal geodesic pathfinding selection based on Dijkstra's algorithm for the characterization of the facial nerve. A total of 119 clinical CT images from preoperative patients have been used to develop this pipeline that produces accurate characterizations of these nerves in the cochlear region and provides reliable measurements for computer-aided diagnosis and surgery planning.

Keywords: Cochlear implant · Deep reinforcement learning · Cochlear nerve · Facial nerve · Surgery planning

1 Introduction

Locating the facial nerve (FN) is crucial for trajectory planning to cochlear implant (CI) surgery. It is also believed that proximity between the FN and the basal turn of the cochlea is associated with an increase incidence in cases of FN stimulation [10] because of the FN's proximity to some of the CI electrodes located in the cochlear structure. FN stimulation can lead to severe involuntary motion and pain, and knowing the FN location when passing by the cochlea structure could help predict cases of FN stimulation and/or adjust cochlear implant stimulation parameters to

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12904, pp. 519–528, 2021.
https://doi.org/10.1007/978-3-030-87202-1_50

mitigate FN stimulation after implantation. The cochlear nerve (CN) is primarily responsible for transmitting the electrical impulses generated in the cochlea to the brain. These impulses will be integrated for hearing and localization of sound. Its diameter is a critical measurement for the diagnosis of inner ear conditions as well as a prerequisite for the successful outcome of a CI surgery [23]. We propose an automated pipeline for the characterization of the CN and FN in CT scans.

Locating thin neural structures is a very challenging task, especially in CT scans. To the best of our knowledge, there is no previous approach to automate the characterization of the CN in CT scans. There are different approaches for FN characterization in CTs most of which focus on surgery trajectory planning, which sometimes dismisses the labyrinth segment shown in Fig. 1. Most of the literature in this field discusses atlas-based segmentation [6,19,20], the use of deformable models [22], or a combination of both [15]. Some of these methods are semi-automatic, based on the manual location of landmarks in precise locations of the FN as seen in [6,22]. Recent approaches use state of the art deep learning methods to segment the FN, as seen in [5,14]. These studies are focused on segmentation for trajectory planning of the CI surgery, and thus focusing on other related structures of the region. However, we can in both studies observe how the performance of the method drops significantly when evaluating the FN.

The first stage of our pipeline is automatic location of landmarks which are designed to provide crucial information for the initialization of the subsequent characterization methods. Annotating 3D landmarks in an unlabelled dataset (as ours) is a much faster process than the full 3D segmentation annotation which is time-consuming and, for these structures, challenging even for experts. Landmark location in medical images is a very active field of research. Initially machine learning approaches were based on regression and classification of hand-crafted features but with the rise of deep learning, various neural networks have been used to automate the location of landmarks as in [17,27].

Analyzing the landmark location problem from an object search prespective, the exhaustive scanning or mapping approach can result in a loss of accuracy or very high computational times. Deep reinforcement learning finds search strategies for locating different structures based on the image information at multiple scales which benefits from the different representations of the image. This resembles the human approach to locating a certain landmark and in this rather challenging scenario, it was found to be the best approach. Ghesu *et al.* in 2016 [8] first used a reinforcement learning agent to navigate through a 3D image. In 2017, Xu *et al.* [25] developed a supervised method to classify actions using image partitioning. The computational cost of these initial approaches was reduced thanks to the patch-based iterative convolutional neural network (CNN) approach proposed in 2018 [12,16] which can be adapted for single or multiple landmark detection. To exploit multi-scale image representation, Ghesu *et al.* extended their reinforcement-learning-based landmark detection in [7,9]. In 2019, Vlontzos A *et al.* [21] proposed the MARL model where multiple agents implicitly communicate by sharing the CNN weights. In 2020, Leroy *et al.* proposed a C-MARL model [11] by adding explicit communication based on sharing

the average weights of the fully connected layers to the MARL model. Communication between agents benefits the system's ability to locate landmarks especially when those present spatial correlation, as in our case of study.

2 Data

The dataset available for this study consists of 119 clinical CT scans from diverse imaging equipment. The CT scans cover the patient's inner ear and were performed before surgery. A region of interest is cropped from each CT scan ($32.1^3\,\mathrm{mm}^3$), with voxel resolution average of 0.3 mm and range of [0.13,0.45]. The dataset is therefore representative of routine clinical scans in terms of anatomy, imaging modality, and image quality. The dataset presents a large variation regarding contrast, intensity and noise levels, which is convinient for the development of this application. All the scans were manually labeled by the main author with the designed landmarks using the *ITK-SNAP* software [26].

The location of seven landmarks has been designed to make each landmark as unique as possible within the structure, so it can be easily differentiated from other anatomical locations. This provides more robustness to the manual annotation which influences the robustness of the model. Furthermore, the placement of these landmarks is directly relevant for assessing various clinical metrics. Because these structures are very small as well as their surrounding structures, this task is very challenging, and the resolution does not allow us to locate very specific features of the nerves. Seven landmarks have been selected, two for the CN and five for the FN. The distribution of the landmarks can be seen in Fig. 1.

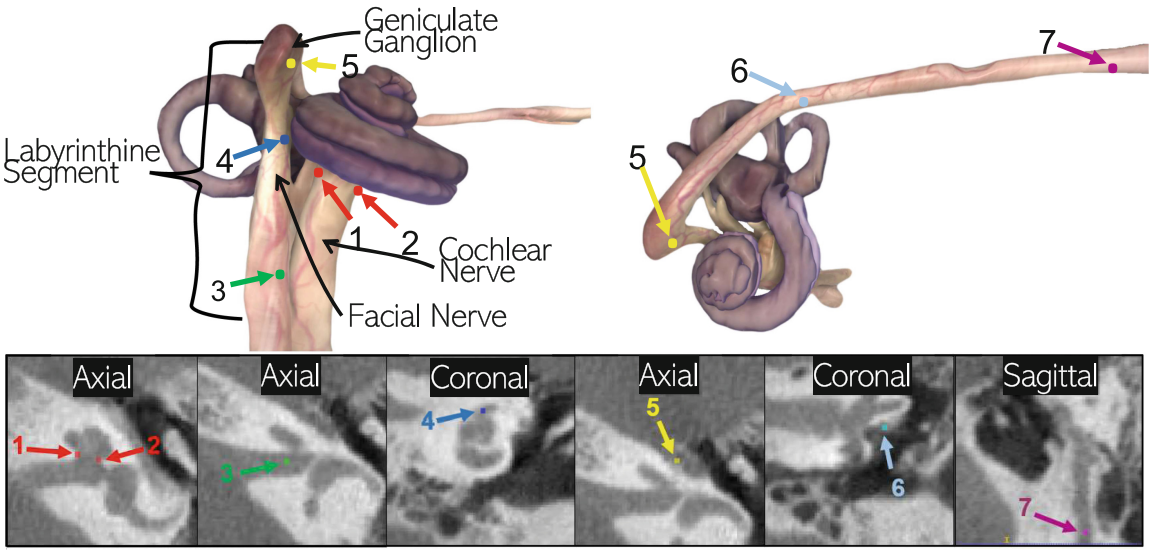


Fig. 1. Top: 3D representation of landmarks in the neural structure. Edited from [18]. Bottom: Landmarks' location in a CT scan (zoomed-in).

3 Methods

Deep-Q-Networks [13] are used to find the optimal strategy for agents to reach their goal. For landmark location, the network architecture resembles the architecture of a typical image classification network, in which convolutional layers extract the relevant features of the state (centered 3D patch in the position of the agent) followed by fully connected layers that map those features to the probability distribution of the Q value of each of the 3D actions (up, down, left, right, forward and backward). Diagonal movements are not considered. The architecture used in this application is shown in Fig. 2 for the case of two agents. The C-MARL model [11] uses multiple agents with implicit (same CNN weights) and explicit communication (average weight sharing of the fully connected layers). We use 21 agents, 3 per landmark, resulting in 21 fully connected networks implemented in the architecture. The dataset containing 119 samples has been randomly split into: 11 samples test set $\approx 10\%$, 12 samples validation set $\approx 10\%$, and 96 training set $\approx 80\%$. The neural networks are trained end-to-end on a 12GB GPU, with a total training time of 6 d. In the training process, the final state is reached when the distance to the landmark is ≤ 1 voxel using the ϵ-greedy search strategy [24]. The forgetting factor γ is set to 0.9 as this has been empirically found to be the best performing option. We use three isotropic resolutions for the multi-scale: 0.9, 0.6, and finally 0.3 mm per voxel dimension. The state of the agent becomes spatially smaller once the agent is oscillating in the current resolution. All the agents must reach the oscillation state in order to either move to the next resolution or to finalize the search. The final model is chosen based on the performance on the validation set.

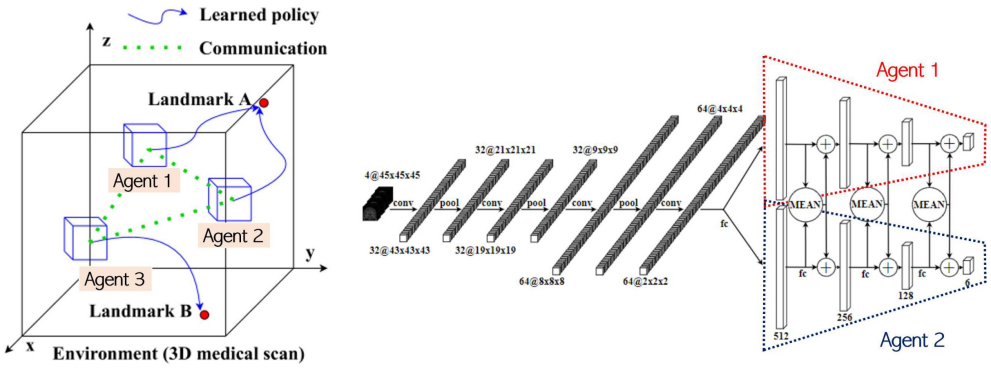


Fig. 2. Left: Diagram of the MARL model Right: architecture of the C-MARL model. Edited from [11,21].

The initial approach for the CN characterization is to measure the distance between landmarks 1 and 2 which are placed on opposite sides of the bony cochlear nerve canal (BCNC) in the axial plane of the CT. To give more robustness to this measurement, the median distance of all the possible distances between the three candidates (one per agent) of each of the landmarks has

been chosen as the final measurement. To characterize this structure further than just the diameter, the area of the cross-section was computed. The cross-sectional slice was thus extracted based on the median location of landmarks 1 and 2 and the known parallelism of the CN and this plane.

The Chan-Vese approach [2] for active contours without edge (ACWE) has been used to segment the cross-section of the CN. The algorithm is based on the mean intensity inside and outside of the curve instead of the gradient, which is convenient for a structure with no clear edges, such as the CN. The region of interest of the cross-section is cropped according to the medial location of the landmarks, and the curve is initialized with a 3-pixel radius circle centered in the centroid of both landmarks. Empirically 15 iterations were found to be sufficient for the algorithm to converge. To fine-tune the regularization parameters of the ACWE algorithm is a tedious task, especially for the big variability of the images intensities and contrast present in clinical CTs. To avoid this task and generate a parameterization of this structure, a least-square ellipse is fitted to the ACWE contour. The CN presents an elliptical shape at this point due to its connection with the spiral shape of the cochlea structure. An example of the processing steps of this method are shown in Fig. 3. To corroborate that the parameterization is robust and reliable, the procedure is extended to multiple slices. This allows us to analyze the continuity in the third dimension and generate a full characterization of the canal. Currently, we select 8 slices to estimate the canal. We use the center of the ellipse fitted in the adjacent slice as the initialization seed for the following one.

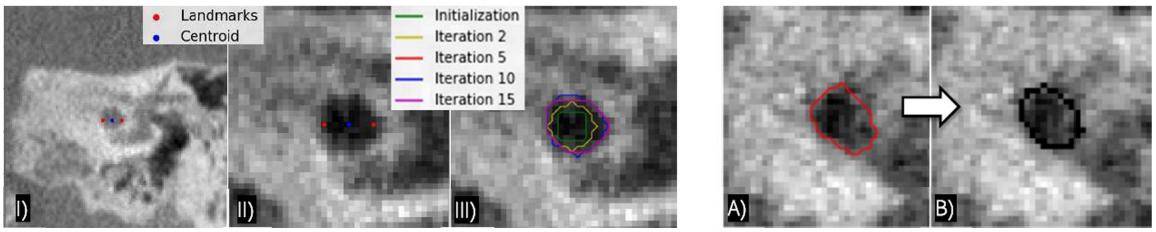


Fig. 3. Left: ACWE segmentations steps (cropping and iterations). Right: Contour of fitted ellipse (B) in black obtained from the ACWE contour (A) in red. (Color figure online)

To characterize the labyrinth segment of the FN, we use Dijkstra's algorithm [3] for geodesic path finding where the edge values are the intensity value of the voxel where the directed graph is pointing. There are three landmarks in the region of interest: number 3 which locates the exit of the FN from the internal acoustic canal, number 4 which locates the closest point of the FN to the cochlea, and number 5 which locates the first geniculate ganglion. To increase the robustness of this approach we design an optimal path selection algorithm. In this algorithm: **I)** Three candidate paths are computed $\text{path}_c^l = \text{path}_6^{3,5}, \text{path}_{18}^{3,5}, \text{path}_{26}^{3,4,5}$, where c represents the connectivity or number of neighbours of each edge and l represents the landmarks used, as illustrated in Fig. 4. **II)** Compute all paths' square derivatives, P_c^l, to have a measurement of the total path intensity variability from adjacent points.

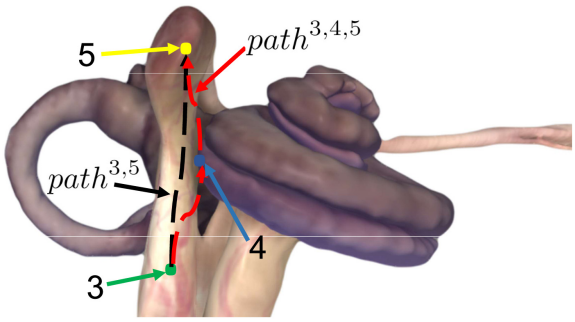


Fig. 4. FN selection algorithm path illustration. Both possibilities of landmarks either 2 (black) or 3 (red) landmarks involved are shown. Edited from [18]. (Color figure online)

$$P_c^l = \sum_{n=1}^{n=N} (\text{path}_c^l[n] - \text{path}_c^l[n-1])^2 \tag{1}$$

III) The path with the minimal square derivative or minimal weight, P_c^l, is chosen as the optimal path.

A 3D representation of the nerve was generated by convolving the binary image containing the binary trace of the FN with a Gaussian kernel with $\sigma = 0.5$ and thresholding the outcome. This produces a normally distributed width of the nerve in the radial direction of the neural structure related to the normal thickness of the nerve in the labyrinth segment (1.23 ± 0.22 mm according to [1]).

4 Results

The results of the described C-MARL model on our test set are shown in Table 1 which includes both a qualitative and quantitative analysis of the performance where the landmark prediction is classified as correct, very close (1 or 2 voxels from optimal position) or wrong.

Figure 5 shows the table with the evaluation of the direct measurement extracted from the landmarks compared to the manual annotation, the signed difference and the absolute error are shown. The average automated BCNC diameters obtained is 2.4mm ($\sigma = 0.31$) which is a value of similar magnitude as the 2.13 mm ($\sigma = 0.44$) found in the clinical study by Fatterpekar *et al.* [4]. The figure also includes an example of the 3D results in a test image.

Figure 6 shows the performance of the path selection algorithm used for the FN tracking as well as an example of the results obtained with the posterior splatting technique and isosurface generation. For all the samples the method selects the best trace of the nerve. Validation and test sets are used to evaluate this method to provide a more significant performance overview as no difference in landmark location performance was found between both data splits.

The results are finally combined in a 3D representation generated with *ITK-SNAP* based on the different segmentations as shown in Fig. 7. The cochlea struc-

Table 1. Mean distance error in mm, μ_{Error}, and the standard deviation, σ_{Error}, for each landmark based on the selection of median location of the 3 candidates generated with the described C-MARL model. Qualitative analysis in percentage of correctly located, very close to optimal location and wrong location (color-labeled).

Landmark	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Overall**
μ_{Error}**[mm]**	0.791	0.561	0.889	0.599	0.988	1.897	2.797	1.218
σ_{Error}	0.321	0.224	0.347	0.435	0.409	0.972	2.140	1.185
Correct [%]	100	90.91	100	100	100	100	81.82	96.1
Very close [%]	0	9.09	0	0	0	0	0	1.3
Wrong [%]	0	0	0	0	0	0	18.18	2.6

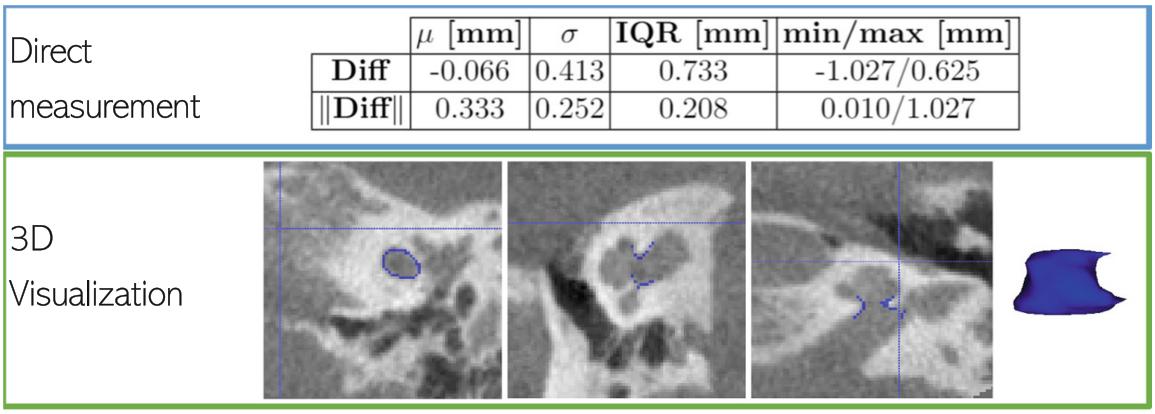


Fig. 5. CN results. Direct measurement: Table showing the performance statistics of the absolute and relative difference between our method and the manual annotations. 3D visualization: CT scan from test set with blue overlay of the 3D ACWE and ellipse fitting approach (left), together with the isosurface 3D representation (right). (Color figure online)

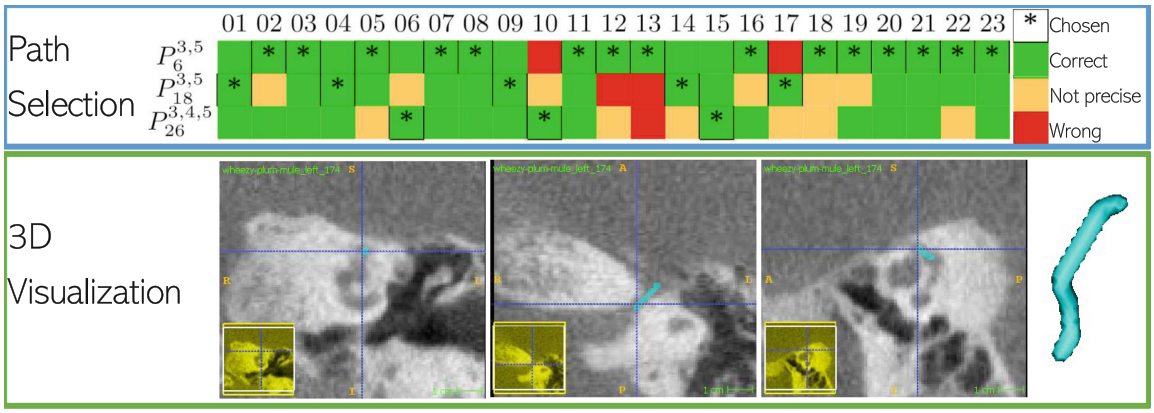


Fig. 6. FN results. Path Selection: Table showing the performance of the selection algorithm in test (1–11) and validation (12–23) scans, all the chosen paths (*) with minimal weight and color-coded the evaluation of the traces. 3D visualization: CT scan from test set with light blue overlay of the splatted selected path (left), together with the isosurface 3D representation (right). (Color figure online)

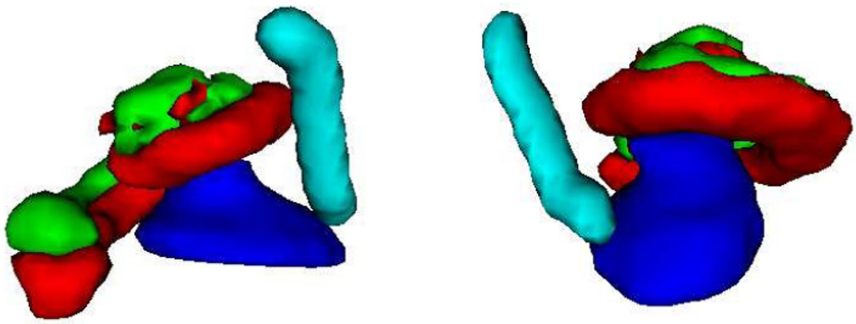


Fig. 7. 3D representation of the segmentation of the anatomical structures. Scala tympani (red), scala vestibuli (green), CN (dark blue), FN (light blue). (Color figure online)

ture has also been included for a better representation of both neural structures together and their relative positions.

5 Discussion and Conclusion

The characterization methods based on the estimated location of the landmarks produce accurate characterization of the neural structures in this region. There is an overall error of 1.218 mm in landmark location within the neural structures, a mean difference of 0.333 mm between the manual and automated BCNC measurement (similar magnitude as the dimension of the voxel), and a correct trace of the FN obtained in all the samples. The results show that both the landmark estimation and the further characterization methods are robust enough to characterize these challenging structures on clinical data. To provide further qualitative analysis of the full characterization with clinical annotations should be obtained.

The main contribution of this study is a novel pipeline using reinforcement learning for the extraction of clinically relevant metrics related to the CN health and the FN anatomy. The automation of the CN measurement in CT has not, to our knowledge, previously been performed. Providing the cross-sectional characterization allows clinicians to study the condition of the CN based on both of its elliptical axes and not only on the axial plane view. This cross-sectional information is rather difficult for non-experts to extract due to the non-orthogonal location relative to the Cartesian axis of the CT image. The correct tracking of the FN without human interaction along the labyrinth segment is a big advantage, as this region is known for the difficulty of its characterization. This pipeline allows for a characterization of its proximity to the cochlear structure in this critical region and helps clinicians prevent the FN stimulation derived from this proximity.

The result of this work is a complete pipeline that automatically segments and parameterizes both neural structures in the close-by area of the cochlea. We have developed an advanced tool for CI surgical planning based on deep learning and computer vision algorithms which may pave the way to support

clinicians in the routine assessment of FN anatomy, related risk assessment of FN stimulation, and the evaluation of CN health.

References

1. Celik, O., Eskiizmir, G., Pabuscu, Y., Ulkumen, B., Toker, G.T.: The role of facial canal diameter in the pathogenesis and grade of bell's palsy: a study by high resolution computed tomography. Brazilian J. Otorhinol. **83**, 261–268 (2017)
2. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Trans. Image Process. **10**(2), 266–277 (2001)
3. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische mathematik **1**(1), 269–271 (1959)
4. Fatterpekar, G.M., Mukherji, S.K., Lin, Y., Alley, J.G., Stone, J.A., Castillo, M.: Normal canals at the fundus of the internal auditory canal: CT evaluation. J. Comput. Assist. Tomogr. **23**, 776–780 (1999)
5. Fauser, J., et al.: Toward an automatic preoperative pipeline for image-guided temporal bone surgery. Int. J. Comput. Assist. Radiol. Surg. **14**(6), 967–976 (2019)
6. Gare, B.M., Hudson, T., Rohani, S.A., Allen, D.G., Agrawal, S.K., Ladak, H.M.: Multi-atlas segmentation of the facial nerve from clinical CT for virtual reality simulators. Int. J. Comput. Assist. Radiol. Surg. **15**, 259–267 (2020)
7. Ghesu, F.C., Georgescu, B., Grbic, S., Maier, A.K., Hornegger, J., Comaniciu, D.: Robust multi-scale anatomical landmark detection in incomplete 3D-CT data. Proc. MICCAI **2017**, 194–202 (2017)
8. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An artificial agent for anatomical landmark detection in medical images. Proc. MICCAI **2016**, 229–237 (2016)
9. Ghesu, F.C., et al.: Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 176–189 (2019)
10. Hatch, J.L., et al.: Can preoperative CT scans be used to predict facial nerve stimulation following CI? Otol. Neurotol. **38**, 1112–1117 (2017)
11. Leroy, G., Rueckert, D., Alansary, A.: Communicative reinforcement learning agents for landmark detection in brain images. In: Kia, S.M., et al. (eds.) MLCN/RNO-AI -2020. LNCS, vol. 12449, pp. 177–186. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66843-3_18
12. Li, Y., et al.: Fast multiple landmark localisation using a patch-based iterative network. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 563–571. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_64
13. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
14. Nikan, S., Osch, K.V., Bartling, M., Allen, D.G., Rohani, S.A., Connors, B., Agrawal, S.K., Ladak, H.M.: PWD-3DNet: a deep learning-based fully-automated segmentation of multiple structures on temporal bone CT scans. IEEE Trans. Image Process. **30**, 739–753 (2021)
15. Noble, J.H., Warren, F.M., Labadie, R.F., Dawant, B.M.: Automatic segmentation of the facial nerve and chorda tympani in CT images using spatially dependent feature values. Med. Phys. **35**, 5375–5384 (2008)

16. Noothout, J.M.H., de Vos, B.D., Wolterink, J.M., Leiner, T., Isgum, I.: CNN-based landmark detection in cardiac CTA scans. CoRR abs/1804.04963 (2018). http://arxiv.org/abs/1804.04963
17. Oktay, O., et al.: Stratified decision forests for accurate anatomical landmark localization in cardiac images. IEEE Trans. Med. Imaging **36**, 332–342 (2017)
18. Trier, P., Karsten Noe, M.S.S.: The visible ear simulator (2020). https://ves.alexandra.dk/
19. Powell, K.A., Kashikar, T., Hittle, B., Stredney, D., Kerwin, T., Wiet, G.J.: Atlas-based segmentation of temporal bone surface structures. Int. J. Comput. Assist. Radiol. Surg. **14**, 1267–1273 (2019)
20. Powell, K.A., Liang, T., Hittle, B., Stredney, D., Kerwin, T., Wiet, G.J.: Atlas-based segmentation of temporal bone anatomy. Int. J. Comput. Assist. Radiol. Surg. **12**, 1937–1944 (2017)
21. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 262–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_29
22. Voormolen, E.H., et al.: Determination of a facial nerve safety zone for navigated temporal bone surgery. Neurosurgery **70**, 50–60 (2012)
23. Waldman, S.D.: Chapter 9 - the vestibulocochlear nerve—cranial nerve viii. In: Waldman, S.D. (ed.) Pain Review, pp. 22–25. W.B. Saunders, Philadelphia (2009). http://www.sciencedirect.com/science/article/pii/B9781416058939000095
24. Watkins, C.J.C.H.: Learning from Delayed Rewards. Ph.D. thesis, King's College, Cambridge, UK, May 1989
25. Xu, Z., et al.: Supervised action classifier: approaching landmark detection as image partitioning. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 338–346. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_39
26. Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006)
27. Zhang, D., Liu, Y., Noble, J.H., Dawant, B.M.: Automatic localization of landmark sets in head CT images with regression forests for image registration initialization. Med. Imaging 2016 Image Process. **9784**, 97841M (2016)

PAPER B

Accurate Localization of Inner Ear Regions of Interests Using Deep Reinforcement Learning

Authors Ana-Teodora Radutoiu, François Patou, Jan Margeta, Rasmus R. Paulsen, and Paula López Diez.

Journal Machine Learning in Medical Imaging. MLMI 2022. Lecture Notes in Computer Science, vol 13583. Springer, Cham.

Year 2022

Status Published

DOI https://doi.org/10.1007/978-3-031-21014-3_43

Accurate Localization of Inner Ear Regions of Interests Using Deep Reinforcement Learning

Ana-Teodora Radutoiu[1(✉)], François Patou[2], Jan Margeta[3], Rasmus R. Paulsen[1], and Paula López Diez[1]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
s194324@student.dtu.dk
[2] Oticon Medical, Research and Technology group, Smørum, Denmark
[3] KardioMe, Research and Development, Nova Dubnica, Slovakia

Abstract. We propose a novel method for automatic ROI extraction. The method is implemented and tested for isolating the inner ear in full head CT scans. Extracting the ROI with high precision is in this case critical for surgical insertion of cochlear implants. Different parameters, such as CT equipment, image quality, anatomical variation, and the subject's head orientation during scanning make robust ROI extraction challenging. We propose to use state-of-the-art communicative multi-agent reinforcement learning to overcome these difficulties. We specify landmarks specifically designed to robustly extract orientation parameters such that all ROIs have the same orientation and include the relevant anatomy across the dataset. 140 full head CT scans were used to develop and test the ROI extraction pipeline. We report an average overall estimated error for landmark localization of 1.07 mm. Extracted ROI presented an intersection over union of 0.84 and a Dice similarity coefficient of 0.91.

Keywords: Region of interest · Deep reinforcement learning · Computed tomography · Inner ear · Landmarks · Orientation

1 Introduction

Automatic region of interest (ROI) detection in medical images is a challenging task as medical images generally present high variability between individuals, scanners, and image acquisition and postprocessing protocols. ROI extraction is a necessary step for almost all medical image analysis pipelines. It is also vital for subsequent image processing tasks that rely on input stability. Accurate ROI extraction can not only improve retrieval efficiency but can also help to classify more easily pathological signs within a reduced region especially when the anatomy is particularly challenging for either the clinicians or the processing software to interpret [13].

Cochlear implants (CI) are used to restore the hearing capacities of patients who suffer from severe to profound hearing loss. CIs are common for infants with

C. Lian et al. (Eds.): MLMI 2022, LNCS 13583, pp. 416–424, 2022.
https://doi.org/10.1007/978-3-031-21014-3_43

congenital deafness. Clinicians evaluate each patient using computed tomography (CT) images, the head orientation is important to assess each case and successfully obtain the relevant measurements for accurate surgical planning. Therefore we had developed an approach to automatically locate this region in clinical full head CT images.

Initially, ROI detection methods were based on bounding boxes from handcrafted features [3,11]. Nowadays, the most common technique is deep learning, where a majority of methods in the literature are designed for 2D medical images [2,13] and other approaches used 2D methods to locate ROIs in 3D [5,8,16]. These methods fail to use the third dimensionality of CT images and need a big amount of annotated data that faithfully represents the anatomical variability and struggle when the anatomy is abnormal. We chose to use a deep reinforcement learning (DRL) based approach that uses landmarks (easier and faster to annotate) to automatically extract the inner ear ROI from CT images. DRL has been successful showing outstanding performance for similar tasks of landmark localization in medical images [1,15]. Other DRL approaches include Navarro et. al. [10], this paper proposes single reinforcement learning agents for organ localization in the torso. They succeeded in effectively finding a ROI around desired organs with a relatively small amount of data. We chose to use a landmark-based approach which we consider will be more robust and could potentially provide more explainability or even help detect abnormalities from an early processing stage [6,13].

2 Data

This study uses 140 full head CT scans from the CQ500 dataset [4] which corresponds to 102 different patients. The dataset has been provided by the Centre for Advanced Research in Imaging, Neurosciences and Genomics (CARING), New Delhi, India [4]. The CT scans are taken from several radiology centers in New Delhi and are collected using various equipment models [4]. All used scans are resampled to have the isotropic voxel spacing 0.5 mm. The image dimensions vary significantly within our dataset. On average the dimensions are $475 \times 475 \times 323$, but the dimensions $x \in [400; 576]$, $y \in [400; 576]$ and $z \in [128; 730]$. All scans are manually labeled with the chosen landmarks using the software 3D Slicer [7]. All the annotations are made public and can be found in https://github.com/AnaTeodoraR/annotations.git.

Choosing relevant landmarks is necessary to characterize the inner ear orientation. The landmarks must be uniquely defined within their structure, so they are easily differentiated from other anatomical points nearby. Eleven landmarks are chosen in total, five assigned for each inner ear ROI and one common for both. The five landmarks for each ROI are the same anatomical points but located on their respective side of the CT scan. All landmarks are associated with a number; Numbers 1–5 are in connection to the right ROI, 6–10 for the left, and 11 is common. Two landmarks are within the inner ear; cochlear apex (nr. 1 and 6) and superior semi-circular canal peak (nr. 4 and 9). Additionally, two landmarks in the cochlea nerve; the midpoint below the base of the cochlea

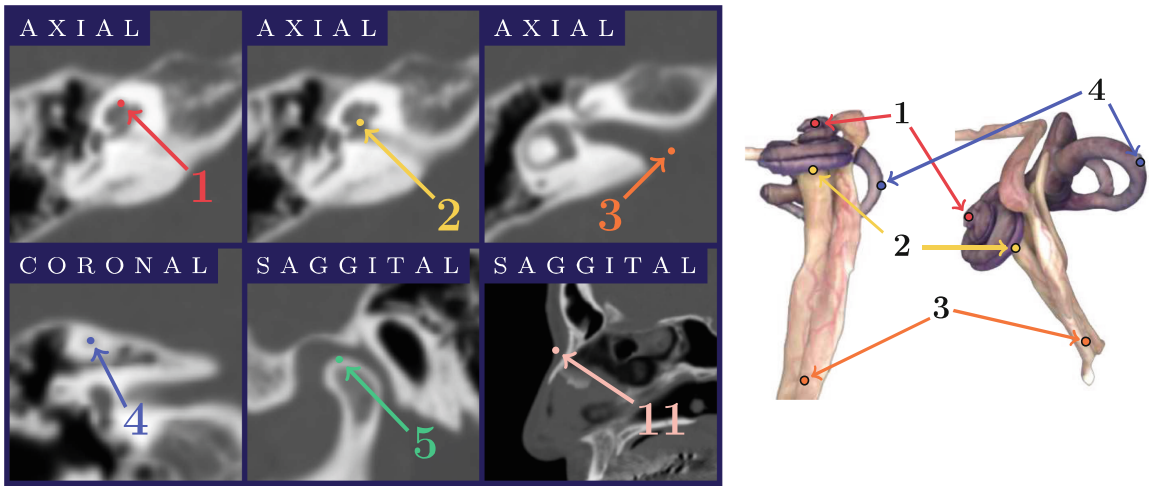


Fig. 1. Landmarks 1–5 and 11 are shown on CT scan nr. 6 and additionally landmarks 1–4 are shown on a 3D representation of an inner ear (edited from [14]). Landmarks 1–5 are for the right inner ear ROI, but the corresponding landmarks 6–10 are placed at the same anatomical points on the left side of the head.

(nr. 2 and 7) and of both the CN and FN further down (nr. 3 and 8). To aid in obtaining a similar orientation the remaining landmarks are; At the top of the condyle in the mandible where the temporomandibular joint starts (nr. 5 and 10) and lastly the nasion (nr. 11) (Fig. 1).

3 Methods

The strategy used for detecting the landmarks is DRL specifically the communicative multi-agent reinforcement learning (C-MARL) model proposed by Leroy et al. [9]. This model uses DRL with multiple agents that communicate and are based on a deep Q-learning network (DQN). The environment is defined as the entire 3D medical scan while an agent is moved voxel by voxel. A DQN is used to predict the optimal Q values given a certain state. As input the network takes states - a 3D patch centered around the agent - and outputs the Q values for each of the possible actions - 3D movement (up, down, left, right, forward, and backward) [9]. The architecture for two agents is shown in Fig. 2. All the agents share the weights of the convolutional layers (implicit communication) while each agent has its own fully connected (FC) layers only sharing the average output from each FC layer (explicit communication) [9]. The agents use multi-scale which enables them to have a spatially higher view of the image in its state. In our implementation, the agents used four scales including the isotropic voxel spacings 3, 2, 1, and 0.5 mm for the patch to represent the state. Initially, the agents will observe the states with the coarsest spacing, but when they start oscillating the spacing will decrease to the next, finer, resolution. The final model uses 11 agents, one per landmark, thus resulting in 11 FC layers and a discount rate of 0.8. The data is divided into training, validation, and test sets each being 112, 14, and 14 CT scans respectively (≈80%, ≈10%, and ≈10%).

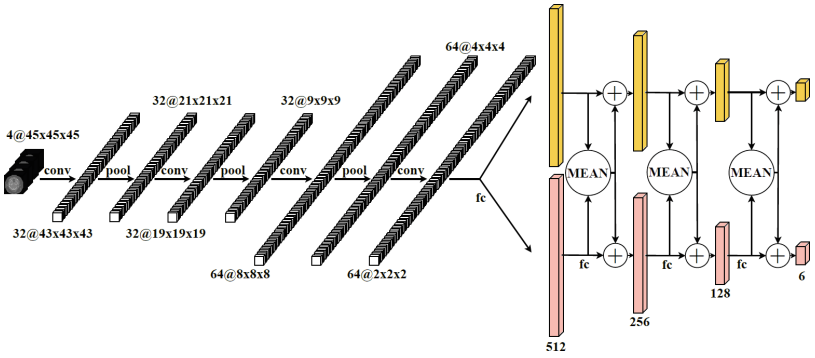


Fig. 2. Visualization of the CNN architecture, when using two agents. The fully connected layers for each agent are colored differently. Edited from [9].

The model was trained on a 12 GB GPU for five days. The training uses an ϵ-greedy strategy ($\epsilon \in [0.1; 1[$). The model which has the best performance on the validation set is further used for testing.

The aim of this study is to extract the two ROIs with similar orientation in a two-step approach. The first step aims to rotate the images to align all heads. The landmarks are used to define the relevant rotation angles that characterize the head orientation. The second step is a regular axis-aligned crop of the inner ear on the newly rotated image. To rotate a 3D medical image, three rotation angles denoted α, β, and γ are used, also known as yaw, roll, and pitch respectively. They describe the head movement in 3D as shown in Fig. 3.

The rotation angles α and β have the potential to be estimated based on the assumption that the corresponding left and right anatomical structures are symmetrical as is the case for the ear anatomy [12]. Technically, α and β are found by projecting the line between the same two anatomical points onto the axial plane and coronal plane respectively, and finding the deviation from the horizontal line, see Fig. 3a and 3b. Four independent estimations of these angles are found using the landmark pairs (1, 6), (2, 7), (3, 8), and (5, 10) and a median of these estimations is applied as the final rotation angle. Likewise, γ is found assuming landmark 5 or 10 and landmark 11 are at an angle of $\theta = 20°$ when projected on the sagittal plane as seen in Fig. 3c. Humans are anatomically different but we estimate 20° to be a good estimation of this anatomy. Finally, γ is defined as the angle which corrects the difference between θ and the estimated ones using the landmarks (Fig. 3c). Once more two different γ values can be estimated, so a median of these will be used as the final γ.

After obtaining the three angles for a single 3D medical scan, the image is rotated accordingly using three individual rotation transformations. The ROI is extracted on the rotated image as a 3D axis-aligned crop from a center point with a customized size. The center point is set to the middle point between two landmarks (landmarks 1 and 2 for the right, 6 and 7 for the left). The radii, r_x, r_y and r_z, of the 3D box, are found as the furthest distance to a set of landmark's respective x, y, and z coordinates (landmarks 1, 2, 3, and 4 for the right ROI,

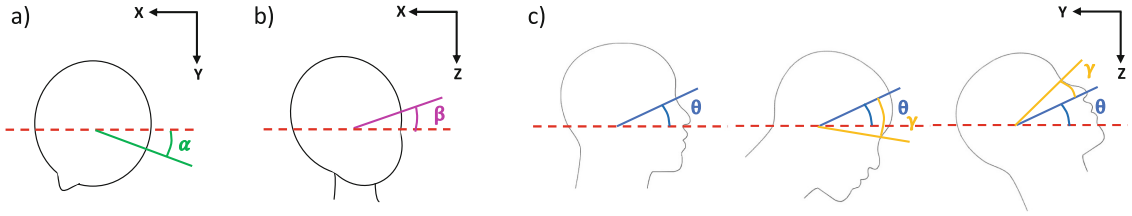


Fig. 3. Sketch of α, β, and γ where the rotation angles have been exaggerated. The angle $\theta = 20°$ describes the desired head position for γ. a) Axial view b) Coronal view c) Sagittal view and γ for the two different orientations.

while 6, 7, 8, and 9 for the left). Moreover, $r_x = r_y = \max(r_x, r_y)$, which results in the ROIs having quadratic axial slices.

A robustness analysis is made by applying a grid of different rotations to an image. The chosen image is the one where the sum of absolute calculated angles (from the annotations) is the smallest, hence the image closest to the reference pose. This image has the estimated angles; $\alpha = 0.19°$, $\beta = 1.08°$, $\gamma = 0.70°$, hence its sum is 1.97°. For our robustness analysis, we exhaustively sampled all 3D rotations on a uniform grid: $\alpha', \beta' \in \{-15, -10, -5, 0, 5, 10, 15\}$ and $\gamma' \in \{-20, -10, 0, 10, 20, 30\}$, where α', β', and γ' are the applied rotation along the third, second and first axis respectively. This results in $7 \cdot 7 \cdot 6 = 294$ manually rotated images. For all generated images their landmarks will be predicted using the C-MARL model. Furthermore, the angle calculation method is applied using the predicted landmarks and compared to the applied rotation. Consequently, the applied rotation will have the opposite sign of the predicted (if predicted correctly).

4 Results

Different C-MARL models have been trained and it was empirically found that using a discount rate of 0.8 and a single agent per landmark, performed best. Using multiple agents per landmark showed only to have a relevant influence on detecting landmark 11. The results of the final model on the test set can be seen in Table 1. Observing the table, the model performs particularly well at detecting landmarks within the inner ear and cochlear nerve (landmarks 1, 2, 3, 4, 6, 7, 8, and 9). For these eight landmarks, the mean distance error is below one millimeter. Depending on the landmark, the estimated errors vary between 0.79–2.11 mm.

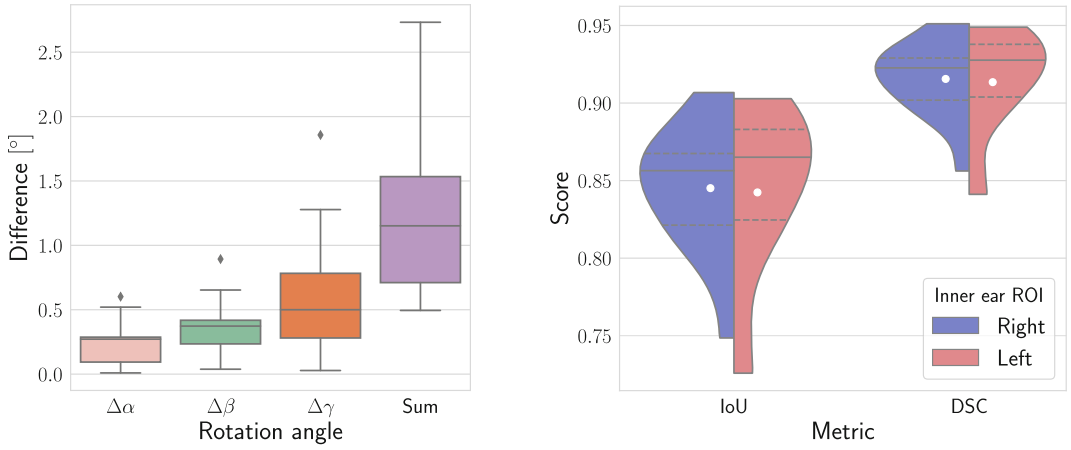
Figure 4a shows box-plots of the rotation angle differences across the test images. Here the rotation angle difference is between the predicted angles and the ones found from the landmark annotations. The maximum deviation of the predicted angle from the estimated (calculated from the annotated landmarks) for α and β is below 1° and the difference for γ is capped at 1.8°, with its upper quartile is below 1°. Figure 4a additionally illustrates a box-plot of the sum of the three differences in a test image which ranges from 0.36°–2.68°.

Now we evaluate the ROI extraction performance. Figure 4b illustrates the intersection over union (IoU) and Dice similarity coefficient (DSC) for both ROIs, additionally, we show the distribution between right and left ROIs. The IoU and DSC are found by comparison with an estimated ROI using the annotated landmarks. Observing Fig. 4b, a small difference between the right and left ROI's prediction exists. The scores for the right and left ROI appear unimodal having a single peak.

The results of the robustness analysis are shown in Fig. 5. Observing the two figures, each 3D point represents a manually rotated image whose three applied rotation angles (α', β', and γ') are its x, y, and z coordinates. The points are color-coded depending on the performance. Looking at Fig. 5a, many of the rotated images have an estimated error of 1–2 mm (purple color). A general tendency of more purple colors for smaller and negative γ' is observed and a decline in performance is seen as γ' increases. The lowest accuracy tends to be gathered at highly negative α', positive β', and positive γ' rotations. Similar patterns can be observed in the corresponding figure for the angle error (Fig. 5b). The error here is the sum of differences between the predicted rotation angles

Table 1. The estimated error (mean distance error, d_{Error}) and standard deviation (σ_{Error}) on the test set. The model predicts the landmarks three times on each test image and a median is used as the final prediction. The last row shows the percentage of detections below one millimeter.

Landmark	1	2	3	4	5	6	7	8	9	10	11	Overall
mean(d_{Error}) [mm]	0.84	0.79	0.87	0.87	1.63	0.83	0.85	0.82	0.99	2.11	1.24	1.07
σ_{Error} [mm]	0.38	0.42	0.64	0.44	1.14	0.28	0.39	0.56	0.31	1.04	0.64	0.75
<1 mm [%]	64.3	78.6	78.6	64.3	35.7	78.6	57.1	85.7	57.1	7.1	50	–



(a) Box-plots of angle differences. (b) Violin-plots of ROI scores.

Fig. 4. Left: Box-plot of the angle differences on the test images (both the individual ones and their sum in an image). **Right:** Violin-plot of IoU and DSC on the test images. The line represents the median, the stripes the upper and lower quartile, and the white dot the mean.

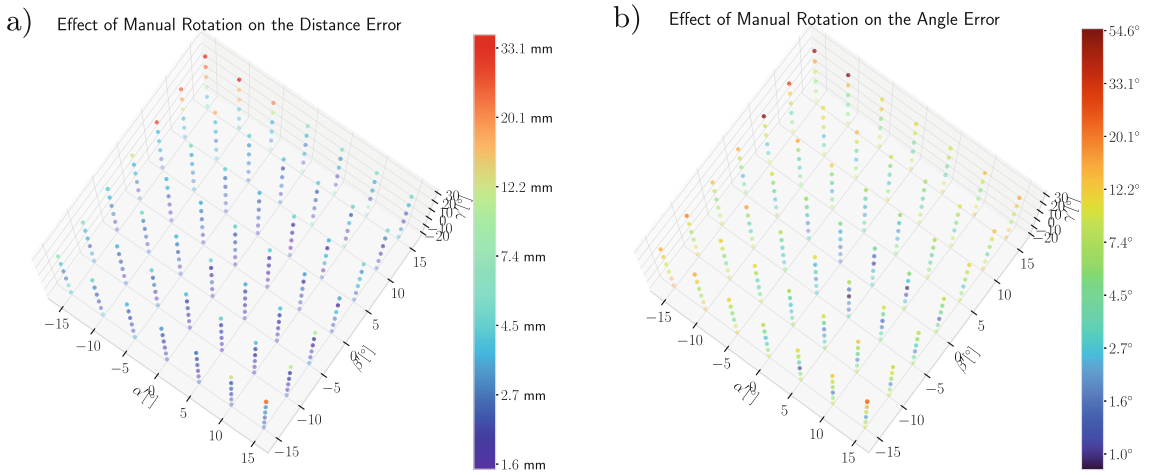


Fig. 5. Each point represents an image that has been manually rotated, the rotation angles used are illustrated as its x, y, and z coordinates. The colors illustrate the estimated error for a landmark (right) and the sum of the angle prediction errors (left). The coloring has a logarithmic scale. (Color figure online)

and the applied (with opposite sign). Only a few rotated images have an angle error below 2° (blue and black colors), while the general error tends to be around 4–10° (green, lime, and yellow colors).

5 Discussion

We found the landmark accuracy shown in this work sufficient for the ROI extraction task. Studying the performance of the individual landmarks, a few points can be made. The agents looking for landmarks 1–4 (and 6–9) are searching close by one another, and potentially benefiting more from the communication. Thus, a reason as to why the highest precision for landmark localization is achieved for these eight landmarks. Interestingly, landmark 11 is isolated from the rest, hence why the single agent has more trouble detecting it. Landmarks 5 and 10 emerge as the most difficult to locate. These landmarks mark the peak of the condyle where the jaw joint starts. The peak is a slightly curved surface, moreover, if a patient has a rotating head when taking the CT scan, other parts of the mandible become visible at the same axial slice. These matters make landmarks 5 and 10 difficult to place, both for the annotator and agents.

Multi-scale is of importance [1], especially if agents are looking for a small structure within a big image. It is important that the agents quickly get an idea of which region the landmark is in. The multi-scale enables the agents to view a larger section of the image and quickly narrow down which region their landmark is located. We implemented 4 levels of multi-scale that were sufficient for our task. The preciseness of the landmark localization influences the angle prediction. Moreover, γ is predicted using the landmarks with the lowest localization precision (landmarks 5, 10, and 11), why this rotation angle has the lowest prediction performance of the three (Fig. 4a).

Regarding the robustness analysis in Fig. 5, negative γ' corresponds to rotating the head downwards in the sagittal plane. This is a more common position of the head, why the model performs better at these rotations. Additionally, $\pm 10°$ with α' and β' further worsens the performance, however, it is also less common for a patient to have these rotations when taking a CT scan. Thus, the results of the robustness analysis reveal that rotations that mirror usual head positions perform best. Data augmentation could potentially help the model to be more robust with extreme head rotations.

Considering the ROI extraction, the overall predictions have on average an IoU score of 0.84 and DSC of 0.91. Since the IoU penalizes false positives and false negatives more than the DSC, it explains the lower scores observed for the IoU. Additionally, the regions are not all fixed to be the same size, so it might be the case that a predicted ROI is contained within an estimated (or oppositely). These cases result in a lower IoU and DSC. We compared our ROI extraction method with another state of the art method as seen in Table 2. Table 2 compares the highest average IoU (the liver) achieved by Navarro et. al. [10] with the left and right ROI results. Both inner ear ROIs have on average a higher IoU than the liver ROI. Our presented method is specifically designed to extract the inner ears with likewise orientation. On the other hand, Navarro et. al. [10] strives to achieve a general axis-aligned ROI detection framework (using DRL) for organs in the torso.

Table 2. Comparing the average IoU for the left and right ROI with the average IoU for the liver (best organ result) from Navarro et. al.

	Navarro et al. [10]	This model right ROI	This model left ROI
Avg IoU	0.80	0.836	0.835

6 Conclusion

This study successfully used a DRL framework for landmark detection in full head CT scans, and utilize it for ROI extraction of the inner ears. Landmarks were localized with an estimated error between 0.78–2.11 mm (on average 1.07 mm) within this difficult anatomical structure. The defined rotation angles gave the ROIs the desired orientation. Two ROIs were extracted from the detected landmarks with an overall average IoU of 0.84 and DSC of 0.91. The method outperforms other DRL approaches for ROI detection as is the proposed by Navarro et. al. [10]. Through this study, we explored the capability of implementing C-MARL for predicting fine structures in full head CT scans. This paves the way for analysis of inner ear ROIs for surgical use.

References

1. Alansary, A., et al.: Evaluating reinforcement learning agents for anatomical landmark detection. Med. Image Anal. **53**, 156–164 (2019). https://doi.org/10.1016/j.media.2019.02.007

2. Bi, L., Kim, J., Kumar, A., Fulham, M., Feng, D.: Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation. Vis. Comput. **33**(6), 1061–1071 (2017)
3. Campadelli, P., Casiraghi, E., Esposito, A.: Liver segmentation from computed tomography scans: a survey and a new algorithm. Artif. Intell. Med. **45**(2–3), 185–196 (2009)
4. Chilamkurthy, S., et al.: Development and validation of deep learning algorithms for detection of critical findings in head CT scans (2018). https://doi.org/10.48550/ARXIV.1803.05854, dataset. http://headctstudy.qure.ai/dataset
5. De Vos, B.D., Wolterink, J.M., De Jong, P.A., Viergever, M.A., Išgum, I.: 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. In: Medical imaging 2016: Image processing, vol. 9784, pp. 517–523. SPIE (2016)
6. Diez, P.L., et al.: Deep reinforcement learning for detection of abnormal anatomies. In: Proceedings of the Northern Lights Deep Learning Workshop, vol. 3 (2022)
7. Fedorov, A., et al.: 3D slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012). https://doi.org/10.1016/j.mri.2012.05.001
8. Hiraman, A., Viriri, S., Gwetu, M.: Efficient region of interest detection for liver segmentation using 3D CT scans. In: 2019 Conference on Information Communications Technology and Society (ICTAS), pp. 1–6 (2019). https://doi.org/10.1109/ICTAS.2019.8703625
9. Leroy, G., Rueckert, D., Alansary, A.: Communicative reinforcement learning agents for landmark detection in brain images. In: Kia, S.M., et al. (eds.) MLCN/RNO-AI -2020. LNCS, vol. 12449, pp. 177–186. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66843-3_18. https://arxiv.org/abs/2008.08055
10. Navarro, F., Sekuboyina, A., Waldmannstetter, D., Peeken, J.C., Combs, S.E., Menze, B.H.: Deep reinforcement learning for organ localization in ct. In: Arbel, T., et al. (eds.) Proceedings of the Third Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 121, pp. 544–554. PMLR, 06–08 July 2020. https://proceedings.mlr.press/v121/navarro20a.html
11. Peng, J., Hu, P., Lu, F., Peng, Z., Kong, D., Zhang, H.: 3D liver segmentation using multiple region appearances and graph cuts. Med. Phys. **42**(12), 6840–6852 (2015)
12. Reda, F.A., McRackan, T.R., Labadie, R.F., Dawant, B.M., Noble, J.H.: Automatic segmentation of intra-cochlear anatomy in post-implantation CT of unilateral cochlear implant recipients. Med. Image Anal. **18**(3), 605–615 (2014). https://doi.org/10.1016/j.media.2014.02.001
13. Sudha, S., Jayanthi, K., Rajasekaran, C., Sunder, T.: Segmentation of RoI in medical images using CNN-a comparative study. In: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), pp. 767–771. IEEE (2019)
14. Trier, P., Noe, K.: The visible ear simulator (2020). https://ves.alexandra.dk/
15. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 262–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_29
16. Ying, W., Cunxi, C., Tong, J., Xinhe, X.: Segmentation of regions of interest in lung CT images based on 2-D Otsu optimized by genetic algorithm. In: 2009 Chinese Control and Decision Conference, pp. 5185–5189. IEEE (2009)

PAPER C

Deep Reinforcement Learning for Detection of Abnormal Anatomies

Authors Paula López Diez, Kristine Aavild Juhl, Josefine Vilsbøll Sundgaard, Hassan Diab,Jan Margeta, François Patou, and Rasmus R. Paulsen.

Journal Proceedings of the Northern Lights Deep Learning Workshop 2022. Vol. 3 (2022)

Year 2022

Status Published

DOI https://doi.org/10.7557/18.6280

Deep Reinforcement Learning for Detection of Abnormal Anatomies

Paula López Diez *[1], Kristine Aavild Juhl[1], Josefine Vilsbøll Sundgaard[1], Hassan Diab[4], Jan Margeta[3], François Patou[2], and Rasmus R. Paulsen[1]

[1]DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
[2]Oticon Medical, Research & Technology group, Smørum, Denmark
[3]KardioMe, Research & Development, Nova Dubnica, Slovakia
[4]The National Medical Research Center for Otorhinolaryngology of the Federal Medico-Biological Agency of Russia, Moscow, Russia

Abstract

Automatic detection of abnormal anatomies or malformations of different structures of the human body is a challenging task that could provide support for clinicians in their daily practice. Compared to normative anatomies, there is a low presence of anatomical abnormalities in patients, and the great variation within malformations make it challenging to design deep learning frameworks for automatic detection. We propose a framework for anatomical abnormality detection, which benefits from using a deep reinforcement learning model for landmark detection trained in normative data. We detect the abnormalities using the variability between the predicted landmarks configurations in a subspace based on a point distribution model of landmarks using Procrustes shape alignment and principal component analysis projection from normative data. We demonstrate the performance of this implementation on clinical CT (Computed Tomography) scans of the inner ear, and show how synthetically created abnormal cochlea anatomy can be detected using the prediction of five landmarks around the cochlea. Our approach shows a Receiver Operating Characteristics (ROC) Area Under The Curve (AUC) of 0.97, and 96% accuracy for the detection of abnormal anatomy on synthetic data.

*Corresponding Author: plodi@dtu.dk

1 Introduction

The detection of abnormal anatomies in medical images has a key role in accurate diagnosis making. The detection of different types of malformations in the clinic is usually performed by visual inspection of clinical images, which makes the diagnosis and detection very sensible to the practitioner's experience and subjectivity. State-of-the-art deep learning methods have shown high performance for automatic detection of anomalies as presented in [3] using well-known architectures. Similar deep learning approaches are also use in a clinical context for anatomical anomalies as in [6]. However, training such a model requires large amounts of labeled medical data, which is challenging and expensive to acquire. Additionally, these approaches suffer from limited generalizability because training data rarely faithfully represents all possible pathological appearances, as images representing the less common malformations are particularly complicated to acquire [1]. It has been shown how simpler methods, such as principal component analysis (PCA), can be employed for anomaly detection in e.g. physiological measurements [2] [8]. While for detection of abnormal anatomies in image data, most studies employ more advanced deep learning methods due to the increased complexity of the data [1, 11].

In this paper, we use a Communicative Multiple Agent Reinforcement Learning (C-MARL) model [9] for landmark detection within the structures of the inner ear. Automatic landmark de-

https://doi.org/10.7557/18.6280

tection is a very active field of research, and with the popularization of deep learning, multiple neural networks have been used for landmark localization. From an object search perspective, the problem of locating landmarks in 3D images can result in drops of accuracy and increased processing times due to the unnecessary exhaustive mapping or scanning of the images. Guesu et al. [5] first showed how a deep reinforcement learning approach allows different agents to learn the optimal policy to locate the landmarks' positions using the image information at different scales. This methodology resembles normal human strategy, where the person who is looking for a certain landmark would initially locate the region of the image where the landmark should be, and zoom-in multiple times for a fine-tuned location of the exact point. To benefit from the use of multiple agents, Vlontzos et al. [13] introduced implicit communication between the agents (MARL model), in which the agents share the weights of the convolutional neural network (CNN) layers. Leroy et al. [9] further included explicit communication between the agents, sharing the average weights of the fully connected layers (C-MARL model). This communication scheme allows for a more robust localization of landmarks, especially when they present a spatial correlation across the dataset.

We base our abnormal anatomy detection on the C-MARL model, with the expectation that when a certain structure is not present in the CT scan, the agents of the C-MARL model trained exclusively on normal cases will not converge to a landmark located in the missing or malformed anatomy. We thus exploit the knowledge about the normal anatomies, as this carries implicit information about the abnormal anatomies. The agents will show a lower degree of agreement regarding the final position of the landmark, when presented with an abnormal anatomy. To show this variation, we project the set of estimated landmarks in a space that is configured according to the normal anatomical variations of the landmarks. In order to do so, we align the landmarks of the training set using the Procrustes analysis [7] and use a PCA to decrease the dimensions of the model. We then examine the variability between the agents from the C-MARL model in the PCA space, to determine whether or not the anatomy is normal. The approach is evaluated on a dataset of clinical CT scans of the inner ear which are labelled with the seven landmarks shown in Figure 1. From these images we artificially remove the cochlea structure, thus artificially simulating a CT scan of a patient with cochlear aplasia. It is known, that this malformation in the inner ear is clinically relevant for the diagnosis of hearing loss [12]. Furthermore, the detection of some type of anomaly in the image can be used not only for setting the special cases apart but also as an initial classification that could potentially be the first block of a sub-categorical classification of the abnormalities.

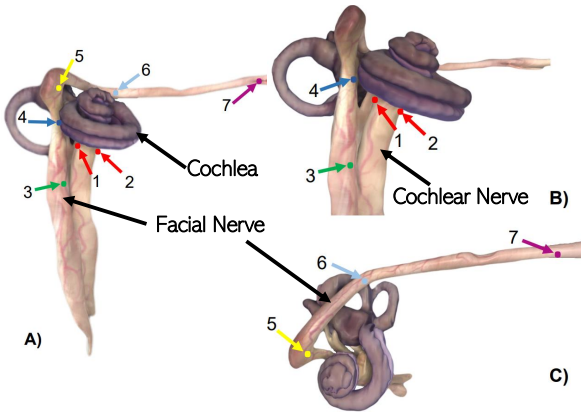


Figure 1: Landmark locations within the facial nerve and cochlear nerve. A) Overview of the seven landmarks B) Close-up of the landmarks 1-4 C) Close-up of landmarks 5-7. Edited from [10]

2 Methods

Our approach is based on using the output variability of a model trained with normative data to detect data anomalies. We employ the C-MARL model [9] for detecting landmarks around the cochlear structure in the inner ear. In this configuration, the deep reinforcement agents navigate through the 3D image (environment) and observe their state, which is defined as a patch of the image centered in the agent location. This patch becomes smaller as the agents get closer to the landmark (multi-scale). Based on the observed state, they take different actions from their action set (move up, down, left, right, forward and backward) and receive a reward, which is a function of the Euclidean distance be-

tween the current position of the agent and the previous one relative to the target point (positive when agent is getting closer and negative otherwise). The expected reward of taking a certain action given a state is known as the Q-value. In deep reinforcement learning the Q-value of a certain state associated with each of the possible actions is estimated by the use of a Deep-Q-Network.

For the C-MARL model the architecture of the Deep-Q-Network resembles a typical image classification architecture, but with a set of fully connected layers for each agent. The model diagram and model architecture is shown in Figure 2. The common CNN weights among all agents provide implicit communication between the agents, meaning the share the same layers responsible of extracting the relevant features for their current state. Meanwhile, the shared average weight of the different fully connected layers allows for implicit communication between agents, sharing information of the layers that are used to map the extracted features from the current state to the Q-value of each action at that point. This setup has been proven specially good when the different landmarks have a consistent spatial correlation as it is the case of the inner ear anatomy. The training configuration employed is the same as presented by López Diez et al. [10], where an overall rate of 2.6% incorrectly located landmarks was reported, with an average error of 1.218 mm.

The agents are randomly initialized within 80% of the image to avoid initialization on an edge. This randomness makes the final estimation of the landmarks a stochastic process. To be able to derive some statistically significant results, we have computed predictions five times for each of the images in the test set.

For normal anatomies, it is assumed that the found landmarks are placed in a certain spatial configuration and that for abnormal cases, the found landmarks will deviate significantly from this configuration. In order to test if a case is within the normal configuration, a point distribution model (PDM) is constructed following the approach in [4]. In the following, the set of found or annotated landmarks from a single scan is called a *shape* and there is point-correspondence over all shapes in the training samples. The shapes are initially aligned using a generalized Procrustes analysis [7]. A similarity transform is used for the alignment and therefore

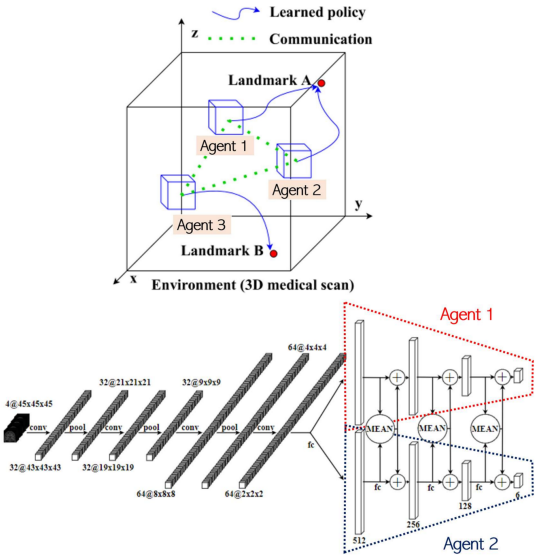


Figure 2: Diagram and architecture of the C-MARL model. Modified from [10]

the PDM will describe the shape variation only and not include the size variation.

Following the Procrustes analysis, a mean shape, $\overline{\mathbf{x}}$ is estimated and the aligned shapes can be used in a principal component analysis (PCA) of shape variation following [4]. The result of the PCA analysis is a set of principal components, concatenated into a matrix Φ, describing the modes of shape variation. A new shape can be synthesized by: $\mathbf{x}_{\text{new}} = \overline{\mathbf{x}} + \Phi\mathbf{b}$. Here $\mathbf{b}$ is a vector of weights controlling the modes of shape variation and Φ contains the first t principal components. A given $\mathbf{x}'$ shape can be aligned to the Procrustes mean and be approximated by the PDM model by projecting the residuals from the average shape into principal component space: $\mathbf{b} = \Phi^T(\mathbf{x}' - \overline{\mathbf{x}})$. The resulting $\mathbf{b}$ vector describes the shape in terms of coordinates in PCA space and is used in the further analysis.

The C-MARL model is trained on images of normal anatomies and their annotated landmarks. These landmarks are also used to build the PDM of the normal anatomy shape configuration. At test time we use three agents per landmark and randomly combine the three predictions into three different shapes of the full anatomical structure. By approximating the shapes with the PDM, we are able to reason about how likely the landmark con-

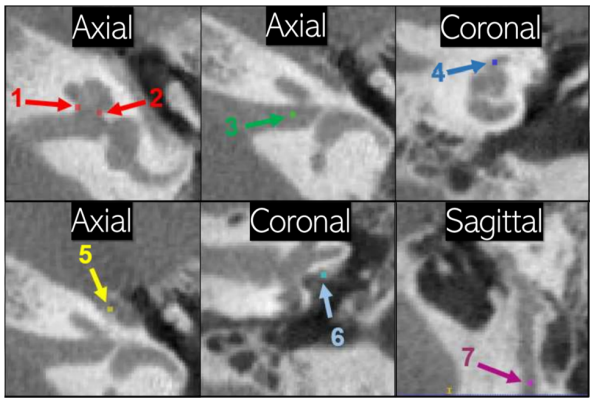


Figure 3: Landmarks' location in a sample CT scan from [10]

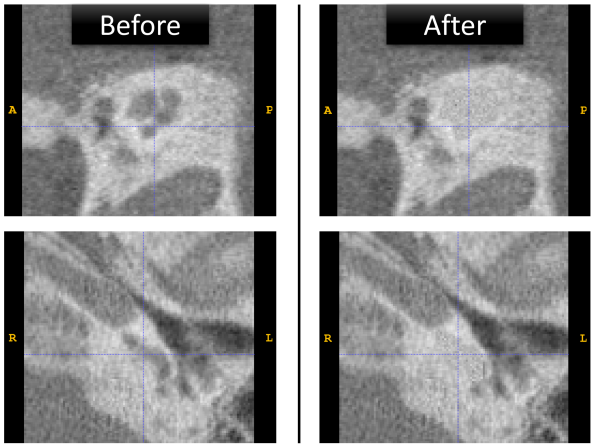


Figure 4: Sample CT scan from test set, before (left) and after (right) the artificial cochlea removal process.

figuration is, as opposed to merely looking at individual distances between single landmarks. Therefore, each shape is projected into the PCA space and described by a vector $\mathbf{b}_i$, where $i = 1, 2, 3$ are the different agents. To quantify the variation between the different agents, we measure the standard deviation between the PCA loadings in the first 6 dimensions (corresponding to 90% of the variation in the PDM).

3 Data

The dataset consists of 120 clinical CT scans of the inner ear from cochlear implant patients with normal inner ear anatomy [10]. From each CT scan, a region of interest is cropped centered in the cochlea with a cubic shape of 32.1x32.1x32.1 mm^3. The average voxel side length of the scans is 0.3 mm in the range of [0.13, 0.45] mm. Each CT scan is annotated with seven landmarks along the facial and the cochlear nerve in the nearby region of the cochlea, as shown in Figure 1 and 3.

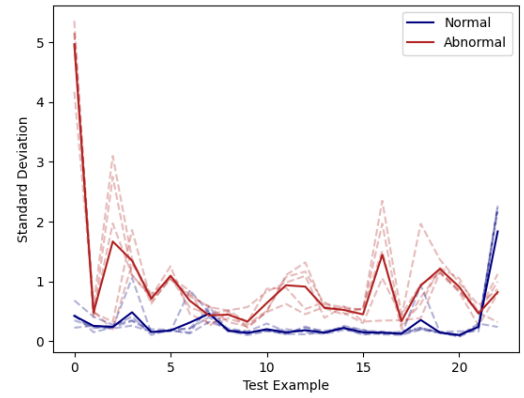
To describe the cochlear image with the landmarks, we use 5 of the 7 landmarks defined in [10]. Landmark one and two are placed on opposite sides of the cochlear nerve in the axial view, number three identifies the point where the facial nerve exits the internal acoustic canal, number four shows the closest point of the facial nerve to the cochlea structure, number five shows the high curvature point of the facial nerve, and six and seven characterize the more elongated part of this nerve in a region more distant from the cochlea. Neither of these landmarks are placed within the cochlea.

For the further analysis for abnormality detection, only the five first landmarks are employed. The first five landmarks are all positioned around the cochlea, while landmarks six and seven are further away from the anatomy in question for this work. The last two landmarks, and thus the six corresponding agents in the C-MARL model, are therefore ignored in the further analysis. The dataset is split into a training set containing 97 CT scans, and a test set containing 23 CT scans. The training data is used to train the C-MARL model for landmark detection, while the test data is used for the rest of the analysis of abnormality detection.

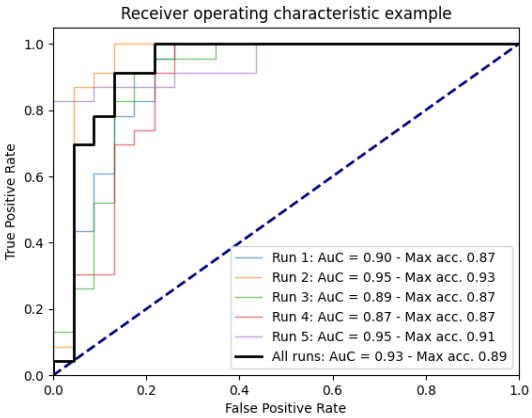
Abnormal CT scans are artificially generated by removing the cochlea structure from the images in the test set, thus generating corresponding pairs of normal and abnormal CT scans with the same surrounding structures of the inner ear. The cochlea is delineated using the ITK-SNAP software [14] to generate a rough segmentation of the cochlear structure. This section of the image is then replaced by Gaussian noise, with mean and standard deviation estimated from the intensities of the surrounding region of the segmentation. An example of the transformation process is shown in Figure 4.

4 Results

For evaluation of the detection of abnormal anatomies, the variation between the PCA loadings for the 15 agents trained for the five landmarks is evaluated. The plot in Figure 6a shows the standard deviation for each corresponding pair of normal and abnormal anatomy, and the graphs show how the standard deviation is increased for the abnormal cases compared to the normal ones. In Figure 6b it can be observed that the ROC curves show a high ROC AUC and maximum accuracy for all five runs, but the averaged method, where all five runs are used to compute the overall standard deviation, shows the highest ROC AUC of 0.97 and a 96% accuracy for detection of abnormal anatomies.

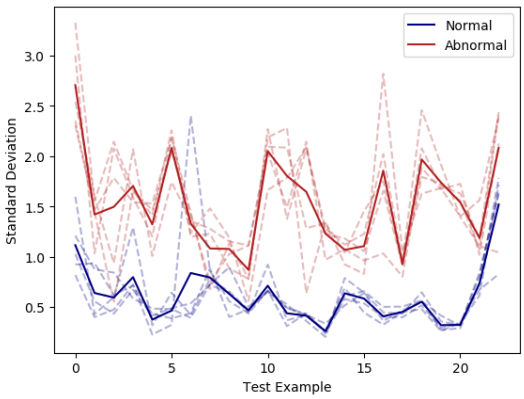


(a)

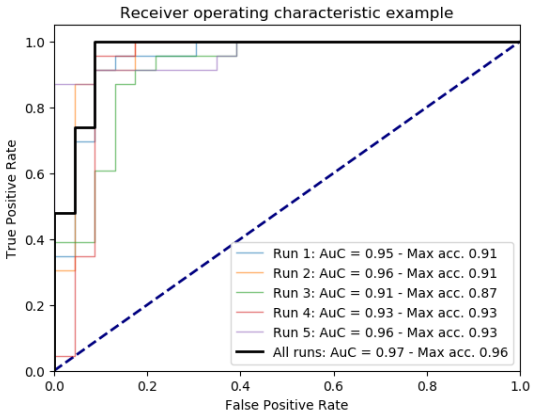


(b)

Figure 5: Results for the direct method on the landmark coordinates. a) Standard deviation between the landmark coordinates for the 15 agents per run (dotted lines) and their average (solid line) for the normal and abnormal cases. b) ROC curves representing the five individual runs (colored lines) and using the average standard deviation across all runs (black line).



(a)



(b)

Figure 6: Results for our proposed PDM and PCA approach. a) Standard deviation between the PCA loadings for the 15 agents per run (dotted lines) and their average (solid line) for the normal and abnormal cases. b) ROC curves representing the five individual runs (colored lines) and using the average standard deviation across all runs (black line).

The results are computed using the test set which contains 23 3D volumes of normative data that has not been seen by the DRL model (trained on 97 images) and the abnormal anatomy version of each image artificially generated as described in Section 3. Overall, the test set contains 46 CT scans: 23 with normal anatomy and 23 with an artificial abnormality. In order to evaluate whether the PCA space of the PDM provides an advantage against analyzing the final configuration of the agents in the original coordinate space, we have also evaluated the variability in the original image space. For each of the three agents per landmark, we calculate the standard deviation of their mutual Euclidean distance. To be able to compare with the representation in the sub-dimensional space, we average all the standard deviations of the five landmarks in each CT scan, so we obtain one value for each scan for each run. Figure 5a shows the representation of these values for the 23 pairs of normal and abnormal CT scans in the test set for each run, together with the overall average among the five runs. It should be noted that the standard deviation in Figures 6 and 5 corresponds to different measurements and the direct comparison should be avoided. The standard deviation seen in Figure 6 is the standard deviation of the distance between each shape projection in the 6-dimensional PCA space of normative data, while in Figure 5 we observe the mean standard deviation of the agents position among the same CT.

Figure 5b shows the ROC curves associated with this approach, as well as the maximum accuracy for each run. In this case, the best performing method is not the average between the five runs, as previously seen. The best results is found for run number two, where the ROC AUC is 0.95, and the maximum accuracy for all runs is 93%.The performance of this method is thus lower than of the proposed method with PDM and PCA, and less stable as the accuracy results of this method have a larger variation between the runs. Figures 6 and 5 cannot be directly compared as in Figure 6 we use the standard deviation of distances in a 6-dimensional PCA space and in 5 we use the standard deviation of distance in mm in the 3-dimensional image space.

5 Discussion

Taking five independent runs of the landmark predictions into account resulted in better results and a more reliable assessment of our method. It can be seen that the accuracy benefits from taking all the runs into account as shown in Figure 6b, where the best performing accuracy of 96% is obtained when taking all five independent runs into consideration. This leads us to believe that the method could benefit from an increased number of runs, given that it appears to provide statistical stability to the method. Potentially, a larger number of agents per landmark could be trained, which may also improve robustness. However, training a greater number of communicating agents would require a significantly longer training and testing time, so a trade-off should be sought between the method's accuracy and the computing cost. We believe our current approach with five runs and three agents per landmark shows a good compromise between accuracy and computational cost.

We have used artificially generated abnormal data to test our approach. This scenario has shown a good performance providing a reliable proof of concept for our method. Further work aims at testing the approach on data with real anomalies and seeing whether or not the results are consistent with the ones presented in this work.

The results show that our proposed approach of measuring the variation of the PCA of the PDM achieves a higher performance compared to measuring the variation of the located landmarks in the image space. The PDM takes into account the full shape of the anatomy, while the inter-agent variance in the image space only takes into account the level of agreement between the multiple agents associated with a certain landmark, but ignores the overall shape. These results show that the method benefits from the shape model, and leads to a more robust prediction of anomalies than just using the analysis in the image space.

Comparison against supervised methods is not possible at this stage due to the lack of labeled data for training, in a further analysis more artificial abnormal data could be generated to train a fully supervised method. However, we consider one of the main advantages of our approach is that it does not require data with abnormalities for the training stages.

6 Conclusion

We have presented an approach for detection of abnormal anatomies in the inner ear based on landmark predictions from a C-MARL deep reinforcement learning model. The method has demonstrated a high performance on the synthetic data for the detection of presence or absence of the cochlea structure. This method manages to both locate the area of the cochlea in the CT scan, and then classify whether or not the structure is present.

References

[1] C. Baur, B. Wiestler, M. Muehlau, C. Zimmer, N. Navab, and S. Albarqouni. Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain mri. *Radiology: Artificial Intelligence*, 3(3):e190169, 2021. doi: 10.1148/ryai.2021190169. URL `https://doi.org/10.1148/ryai.2021190169`.

[2] L. Ben Amor, I. Lahyani, and M. Jmaiel. PCA-based multivariate anomaly detection in mobile healthcare applications. In *Proc. International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*, pages 1–8, 2017. doi: 10.1109/DISTRA.2017.8167682.

[3] R. Chalapathy and S. Chawla. Deep learning for anomaly detection: A survey. 1 2019. URL `http://arxiv.org/abs/1901.03407`.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. doi: https://doi.org/10.1006/cviu.1995.1004.

[5] F. C. Ghesu, B. Georgescu, S. Grbic, A. K. Maier, J. Hornegger, and D. Comaniciu. Robust multi-scale anatomical landmark detection in incomplete 3d-CT data. In *Proc. MICCAI*, pages 194–202, 2017. ISBN 978-3-319-66182-7. doi: 10.1007/978-3-319-66182-7_23.

[6] R. S. Gill, S.-J. Hong, F. Fadaie, B. Caldairou, B. C. Bernhardt, C. Barba, A. Brandt, V. C. Coelho, L. d'Incerti, M. Lenge, M. Semmelroch, F. Bartolomei, F. Cendes, F. Deleo, R. Guerrini, M. Guye, G. Jackson, A. Schulze-Bonhage, T. Mansi, N. Bernasconi, and A. Bernasconi. Deep convolutional networks for automated detection of epileptogenic brain malformations. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Proc. MICCAI*, pages 490–497. Springer, 2018. ISBN 978-3-030-00931-1. doi: 10.1007/978-3-030-00931-1_56.

[7] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/bf02291478.

[8] V. A. Krenn, C. Fornai, N. M. Webb, M. A. Woodert, H. Prosch, and M. Haeusler. The morphological consequences of segmentation anomalies in the human sacrum. *American Journal of Biological Anthropology*, 12 2021. ISSN 2692-7691. doi: 10.1002/ajpa.24466. URL `https://onlinelibrary.wiley.com/doi/10.1002/ajpa.24466`.

[9] G. Leroy, D. Rueckert, and A. Alansary. Communicative reinforcement learning agents for landmark detection in brain images. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, pages 177–186. Springer, 2020. ISBN 978-3-030-66843-3. doi: 10.1007/978-3-030-66843-3_18.

[10] P. López Diez, J. V. Sundgaard, F. Patou, J. Margeta, and R. R. Paulsen. Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. In *Proc. MICCAI*, pages 519–528. Springer, 2021. ISBN 978-3-030-87202-1. doi: 10.1007/978-3-030-87202-1_50.

[11] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, G. Langs, and U. Schmidt-Erfurth. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE Transactions on Medical Imaging*, 39(1):87–98, 2020. doi: 10.1109/TMI.2019.2919951.

[12] L. Sennaroğlu and M. D. Bajin. Classification and current management of inner ear malformations. *Balkan Medical Journal*, 34, 08 2017. doi: 10.4274/balkanmedj.2017.0367.

[13] A. Vlontzos, A. Alansary, K. Kamnitsas, D. Rueckert, and B. Kainz. Multiple Landmark Detection Using Multi-agent Reinforcement Learning. In *Proc. MICCAI*. Springer, 2019. doi: 10.1007/978-3-030-32251-9_29.

[14] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006. doi: 10.1016/j.neuroimage.2006.01.015.

PAPER D

Deep Reinforcement Learning for Detection of Inner Ear Abnormal Anatomy in Computed Tomography

Authors Paula López Diez, Kristine Sørensen, Josefine Vilsbøll Sundgaard, Khassan Diab, Jan Margeta, François Patou, and Rasmus R. Paulsen.

Journal Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, vol 13433. Springer, Cham.

Year 2022

Status Published

DOI https://doi.org/10.1007/978-3-031-16437-8_67

Deep Reinforcement Learning for Detection of Inner Ear Abnormal Anatomy in Computed Tomography

Paula López Diez[1(✉)], Kristine Sørensen[1], Josefine Vilsbøll Sundgaard[1], Khassan Diab[4], Jan Margeta[3], François Patou[2], and Rasmus R. Paulsen[1]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
plodi@dtu.dk
[2] Oticon Medical, Research and Technology group, Smørum, Denmark
[3] KardioMe, Research and Development, Nova Dubnica, Slovakia
[4] Tashkent International Clinic, Tashkent, Uzbekistan

Abstract. Detection of abnormalities within the inner ear is a challenging task that, if automated, could provide support for the diagnosis and clinical management of various otological disorders.Inner ear malformations are rare and present great anatomical variation, which challenges the design of deep learning frameworks to automate their detection. We propose a framework for inner ear abnormality detection, based on a deep reinforcement learning model for landmark detection trained in normative data only. We derive two abnormality measurements: the first is based on the variability of the predicted configuration of the landmarks in a subspace formed by the point distribution model of the normative landmarks using Procrustes shape alignment and Principal Component Analysis projection. The second measurement is based on the distribution of the predicted Q-values of the model for the last ten states before the landmarks are located. We demonstrate an outstanding performance for this implementation on both an artificial (0.96 AUC) and a real clinical CT dataset of various malformations of the inner ear (0.87 AUC). Our approach could potentially be used to solve other complex anomaly detection problems.

Keywords: Deep reinforcement learning · Anomaly detection · Inner ear · Congenital malformation

1 Introduction

Sensorineural hearing loss (SNHL) in children is a major cause of disability. Generally SNHL is detected early in many parts of the world, which allows the prescription of interventions that mitigate the risk of abnormal social, emotional and communicative development. Such interventions include Cochlear Implant (CI) therapy which is prescribed each year to about 80,000 infants and toddlers. Congenital SNHL is sometimes the consequence of an abnormal embryonic development. Resulting malformations are generally classified according to

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13433, pp. 697–706, 2022.
https://doi.org/10.1007/978-3-031-16437-8_67

two categories: membranous malformations, which are not observable in conventional medical scans, and congenital malformations, which can be detected by Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) [4]. These cases raise surgical challenges during surgical planning of the CI therapy and during the surgery itself, often requiring the surgeon to discover and adapt to the anatomy of the malformation during the operation. Anticipating the presence of such malformations from standard imaging modalities is a complex task even for expert clinicians. Categories for these malformations have been described by Sennaroğ L. *et al.* [16], and heuristics have been proposed to help identify them, such as the ones proposed by Dhanasingh A. *et al.* [7]. These heuristics are however of limited use to inexperienced otologists and ear, nose, and throat (ENT) surgeons, who, given the rarity of some of these conditions, cannot easily learn to detect the associated image patterns reliably. We take a first step towards assisting otologists and ENT surgeons in screening or detecting inner ear malformation by proposing the first automated method to detect these anomalies from clinical CT scans.

Different state-of-the-art deep learning methods have shown high performance for automatic detection of anomalies as presented in [5]. Deep learning approaches mostly based in convolutional neural networks used for classification have been used in a clinical context for anatomical anomalies [8]. Training such models requires large amounts of labeled medical data that faithfully represent these anomalies, which is challenging and expensive to acquire, especially because datasets are usually imbalanced because pathological cases are generally rare [17]. We propose a method that is trained uniquely on normative data for landmark location, which makes the approach suitable for adaptation to other anatomies. Knowledge of normal anatomical structural shapes and arrangements acquired during landmark location training brings implicit information for detecting anomalies within that region. Our method is based on multiple landmark location in CT scans of the inner ear. Because we aim to detect abnormalities indirectly by evaluating the output of the model, we define the landmark location as an object search problem and choose to use a deep reinforcement learning (DRL) architecture. We use both the communicative multiple agent reinforcement learning (C-MARL [11]) model and the standard multiple agent reinforcement learning (MARL [19]) model to locate a set of landmarks in the inner ear. We extract two pieces of critical information from these models: First, the variability of the predicted location of a certain landmark across different runs/agents which we evaluate in a subspace defined by the normative data landmarks after they are all aligned using Procrustes, and a principal component analysis (PCA) of the shape variation is performed to define the subspace as presented by López Diez *et al.* in [12]. Second, as a measurement of abnormality, we use the distribution of the predicted Q-values for each agent over the last ten states, including the final position where the landmark is placed. We initially test our approach using a small set of landmarks in a tight crop of the CT images centered on the cochlea versus synthetically generated images of a specific type of inner ear malformation called cochlear aplasia. Furthermore, we

tested the approach in real clinical data using a set of twelve landmarks in a bigger crop of the inner ear.

Simpler methods such as PCA can be employed for anomaly detection in physiological measurements [3,10]. Several groups have used models trained on healthy anatomies to derived the detection of anomalies. While conceptually close to the approach we propose here, the methods have relied on spatial autoencoders or CycleGANs, as described by Baur C. *et al.* [1,2], or segmentation models such as the Bayesian UNet used by Seebock *et al.* [15]. These approaches lack the spatial highlighting and interpretability that our landmark-based approach provides by using highly relevant points of interest defined according to the anatomical malformations.

2 Data

We use two different datasets to test our approach. Our first dataset consists of 119 clinical CT scanners from diverse imaging equipment. These images consist of a region of interest (ROI) with a size of ($32.1^3\,\mathrm{mm}^3$) with the cochlea in its center and an average voxel resolution of 0.3 mm. To test our approach, we synthetically generated abnormal inner ear CT scans from the original images by removing the cochlea (simulating cochlear aplasia) from the images, thus generating corresponding pairs of normal and abnormal CT scans with the same surrounding structures. The cochlea was segmented using ITK-SNAP software [20] and then replaced by Gaussian noise with mean and standard deviation estimated from the intensities of the tissue surrounding the segmentation [12]. An example of the transformation process as well as the location of the anatomical landmarks we use are shown in Fig. 1. This dataset will be called the **Synthetic Set** from now on.

Our second dataset consists of 300 normal anatomy CT scans from heterogeneous sources and 123 CT scans of inner ears that present diverse congenital malformations. This unique dataset contains full-head CT images of CI patients acquired through different CT scanners. This dataset will be referred as the

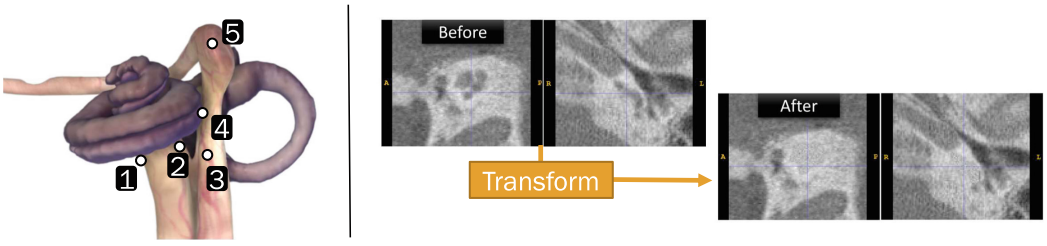


Fig. 1. Left: Set of landmarks used in the Synthetic Set. **1, 2** - Opposite sides of bony cochlear nerve canal in axial view. **3** - Facial Nerve (FN) exiting the Internal Acoustic Canal. **4** - Closest point of FN and cochlea. **5** - Geniculate ganglion of the FN. Edited from [18]. **Right:** Example image of CT scan from test set, before and after the synthetic image generation by inpainting.

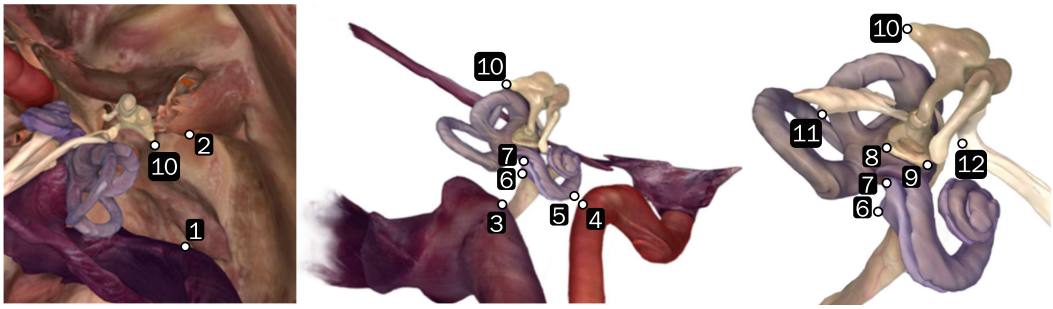


Fig. 2. Set of landmarks annotated in normal anatomy CT images in the Real Abnormality Set. **1** - Sigmoid Sinus (SS) (closest point to EAC). **2** - External Acoustic Canal (EAC) (closest point to SS). **3** - Jugular Bulb (closest point to Round Window (RW)). **4** - Carotid Artery(CA) (closest point to Basal turn of the cochlea). **5** - Basal Turn (closest point to JB). **6, 7** - Anterior and posterior edges of RW. **8, 9** - Anterior and posterior crus of staples. **10** - Short Process of Incus. **11** - Pyramidal Process **12**- Cochleariform Process. Edited from [18].

Real Abnormality Set further on. Out of the 300 normal ears, 175 were manually annotated by an expert with 12 landmarks that define key points of this anatomy. To optimally characterize certain points of interest, these landmarks were designed in close collaboration with our clinical partner, an ENT surgeon specialized in CI therapy in abnormal anatomies. These landmarks are presented in Fig. 2. Simultaneously, the same ROI of 80^3 mm^3 was extracted from the full-head CTs by using the location of the mandible joint and the beginning of the internal acoustic canal for both normal and abnormal anatomies. All images were re-sampled to a 0.5 mm isotropic resolution.

3 Methods

DRL for Landmark Location. Deep-Q-Networks [14] are used to find the optimal strategy for agents to reach their goal. These agents navigate through the 3D image (environment) and observe their state, which is defined as a patch of the image centered on the agent location. This patch becomes smaller as the agent gets closer to the landmark (multi-scale). Based on the observed state, the agent performs one action from the action set (move up, down, left, right, forward, and backward) and receives a reward, which is a function of the Euclidean distance between the current position of the agent and the previous position relative to the target point (positive when agent is getting closer and negative otherwise). The expected reward of taking a certain action given a state is known as the Q-value. In deep reinforcement learning, the Q-value of a certain state associated with each of the possible actions is estimated by the use of a Deep-Q-Network. The architecture of the Deep-Q-Network used for landmark location resembles a typical image classification architecture, but with a set of fully connected layers for each agent. The architecture of the model is shown in Fig. 3. The common convolutional neural network weights among all agents provide implicit

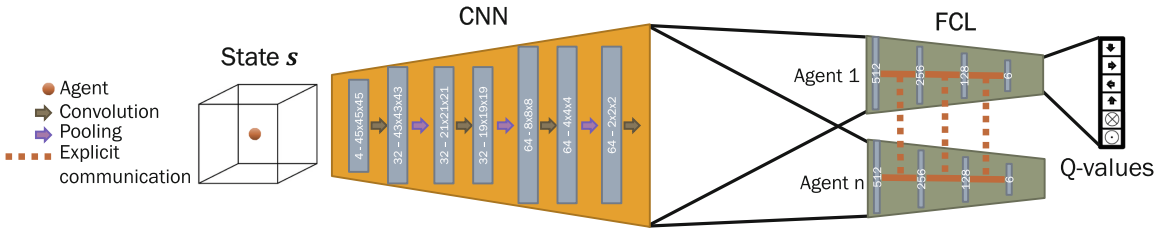


Fig. 3. Diagram of the DRL model used. The explicit communication connections are used in the C-MARL model, but not in the MARL model.

communication between the agents, meaning they share the same layers responsible for extracting the relevant features for their current state. Meanwhile, the shared average weight of the different fully connected layers allows for implicit communication between agents, sharing information of the layers that are used to map the extracted features from the current state to the predicted Q-value of each agent. This setup has been proven especially effective when the different landmarks present a consistent spatial correlation as it is the case with inner ear anatomy [13]. We trained a C-MARL in the normal anatomies of the Synthetic Set and the Real Abnormality Set. Finally we also trained the MARL model on the normal samples of the Real Abnormality Set as we expect that the explicit communication between agents might influence the variability of the model output when facing an abnormal anatomy. The training configuration employed is the same as presented in [13].

PCA Shape Distance Method. The defined landmarks are placed in a spatial configuration which reveals consistency between patients with normal anatomy. We expect that for abnormal cases the landmarks predicted by the model will deviate significantly from this configuration and from one another. In order to test if a case is within the normal configuration, a point distribution model (PDM) is constructed following the approach presented in [6]. We will refer to a full set of landmarks in an image as *shape* where we know there is a point correspondence across all shapes in the training data. The alignment between all the annotated landmarks in normal anatomy is derived using Procrustes analysis [9]. Using this transform, we obtain a PDM that describes the shape variation only and that is invariant to size variation. Once the training shapes are all aligned, a mean shape is computed $\overline{\mathbf{x}}$, followed by a PCA of the shape variation [6]. The outcome of the PCA analysis is a set of principal components concatenated into a matrix Φ, which describes the modes of shape variation. A new shape $\mathbf{x}_{\text{new}}$ can be then defined as: $\mathbf{x}_{\text{new}} = \overline{\mathbf{x}} + \Phi\mathbf{b}$. The vector $\mathbf{b}$ denotes weights controlling the modes of shape variation and Φ contains the first t principal components. We chose to use a t value such that 90% of the shape variability is contained in the Φ matrix. We found $t = 11$ for the 36-dimension space defined by the twelve 3d-landmarks from Fig. 2 over the 175 annotated normal anatomy images of the Real Abnormality Set and $t = 6$ for the 92 normal anatomy images annotated with five landmarks described in Fig. 1 from the Synthetic Set. A given $\mathbf{x}'$ shape can be aligned to the Procrustes mean and be approximated by the PDM model

by projecting the residuals from the average shape into principal component space: $\mathbf{b} = \Phi^T(\mathbf{x}' - \overline{\mathbf{x}})$. The vector $\mathbf{b}$ describes the shape in terms of coordinates in the PCA space. In this space we evaluate the distance between the different shapes predicted by the model. We then define the distance $d_{ji} = ||b_i - b_j||_2$ which measures the variation of all the different shapes predicted for a certain image. Finally we compute the standard deviation of this distribution of distance values for a certain image, D_{image}, which measures the level of agreement among the multiple predictions computed in the PCA space defined by normative shapes. A sketch of this approach is shown in Fig. 4a.

Q-Value History Distribution Method. Using deep Q-learning means that the network is trained to estimate the Q-value, or estimated reward, of taking a certain action given a certain state. Our hypothesis is that if the current state of the agent that is looking for a certain landmark resembles the normal anatomical configuration of such a region, the Q-values will present a uniform distribution, as the agent should not expect a high reward for moving in a certain direction. On the other hand, when the anatomy of the state does not look like what the agent would expect, the Q-values should be less uniformly distributed, pushing the agent to move in a certain direction. To test this hypothesis, we have computed a measurement of the variability within the distribution of the predicted Q-values of the action set. To compute this abnormality measurement we collect the buffer with the predicted Q-values of the last 10 states of the agent, which have empirically be found sufficient to define the later states of the landmark search procedure. These Q-value vectors are then normalized for each agent and merged together with the Q-values of the different runs for each specific landmark. Then, the standard deviation of the Q-values distribution is computed for each landmark u_n. These uncertainty measurements are then joined together into a single value per image $U_{\text{image}} = \sqrt{\sum_n u_n^2}$. An overview of the process is outlined in Fig. 4b.

The combination of both methods has been computed to evaluate its joint performance. Due to the different magnitude of the measurements of each method, a weighting factor has been included in the combination so both methods have a more balanced contribution. The weighting factor w is defined as $w = \frac{\text{median}(D_{\text{training}})}{\text{median}(U_{\text{training}})}$. Then the combination of both methods is defined as $C_{\text{image}} = \sqrt{D_{\text{image}}^2 + (wU_{\text{image}})^2}$ to analyze the joint performance.

4 Results

Each of our tests has been evaluated over five runs for a more rigorous analysis. Both methods described in the previous section where initially tested in the Synthetic Set. The C-MARL model was trained on 92 images with three agents per landmark for the five landmarks shown in Fig. 1. The average error of landmark location is 0.814 mm in the test set. Our method is tested on 27 normal anatomy CT crops and their corresponding artificially created abnormal anatomy scans. The results are shown in Fig. 5a) and d).

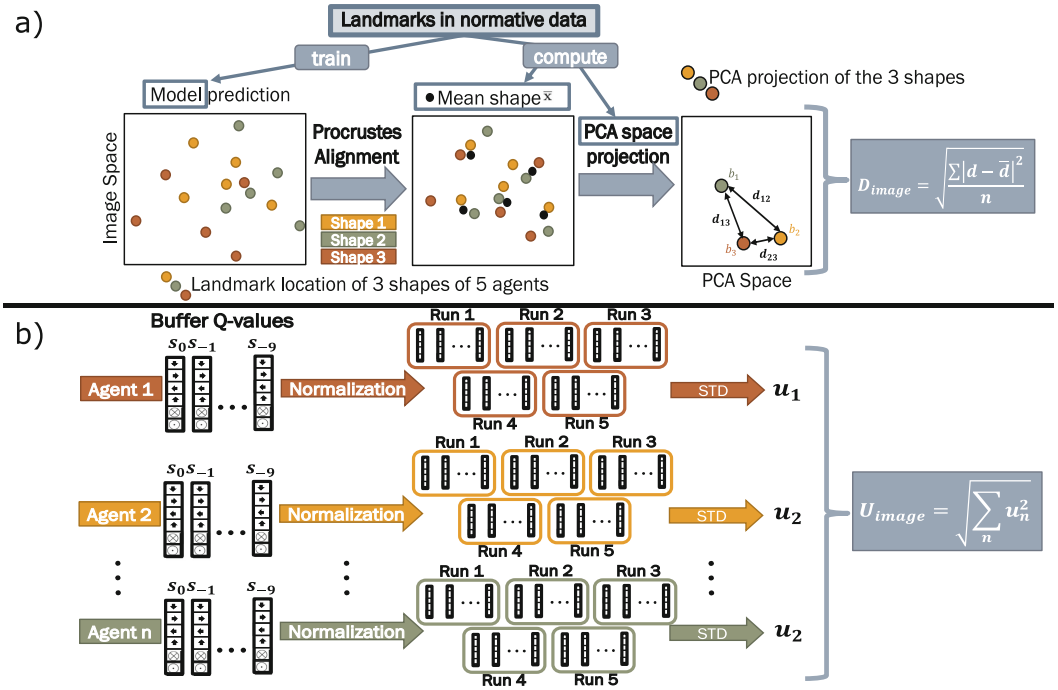


Fig. 4. Diagram of the computation process **a)** D_{image} **b)** U_{image} computation process.

To test the performance in the Real Abnormality Set, we trained a C-MARL model with one agent per landmark for the twelve landmarks described in Fig. 2. The model was trained on 150 CT scans of patients with normal inner ear anatomy with an average test error of 1.74 mm for landmark location over the 25 images of the test set. To test our method we used 123 other different CT scans of normal anatomy and 123 of congenital malformations from the Real Abnormality Set, over five runs. The results of our method are shown in Fig. 5b) and e). As was expected, there is a significant drop in performance when comparing the results in the smaller ROI of the Synthetic Set shown in Fig. 5d) and the results with the same architecture (with a different set of landmarks) shown in Fig. 5e). However, we expected our method would benefit from using the MARL model, which does not include the connections that are responsible for the explicit communication between agents. This means that the agents would not share explicit information about their location and search procedure while looking for the landmarks. This makes agents more independent from each other and less tied to the spatial correlation among them. Our hypothesis is that avoiding this communication will derive greater values for the abnormality measurements when facing an anomaly. The MARL model was trained in the exact same configuration as the C-MARL model and obtained an average error of 1.99 mm on landmark location accuracy. The results of applying our method in this configuration can be observed in Fig. 5c) and f). It can be observed that the method does indeed perform better without explicit communication connections in the model.

The combination of both methods C_{image} shows an improved performance for the Synthetic Set as shown in Fig. 5d), and a very close performance to the best-performing method for the Real Abnormality Set see Fig. 5e) and f). We consider that the combination should be used as a more stable measurement which shows an area under the curve (AUC) of 0.96 for the artificial dataset and 0.86 for the large clinical dataset.

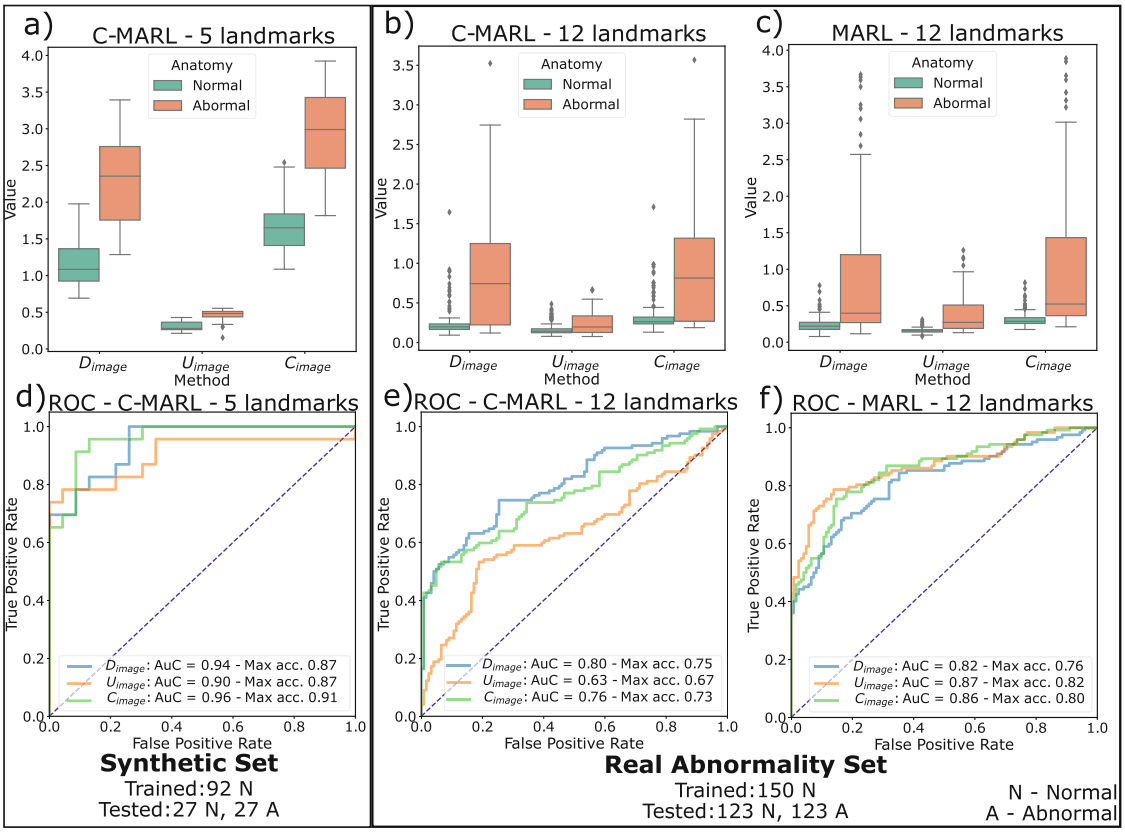


Fig. 5. Evaluation of the different methods over five runs for normal and abnormal anatomies. **Left:** Using 5 landmarks in the Synthetic Set. **Right:** Using 12 landmarks in the Real Abnormality Set. **a), b)** and **c)** Abnormality measurements distribution. **d), e)** and **f)** Derived ROC curves. **a), b), d)** and **e)** C-MARL model. **c)** and **f)** MARL model.

5 Conclusion

We have demonstrated that we can detect abnormal inner ear anatomies by solely training a DRL model on normative data and evaluating the output variability of certain implicit information. This information looks at the relative position of the predicted landmarks over different runs/agents in a subspace defined by the normative annotations as well as the distribution of the Q-values of the last iterations of the agents as a measurement of the uncertainty of the final location. Our MARL model achieved the best performance with an AUC of 0.87 in clinical data which is a high score for such a complex classification problem. We showed

how uncertainty information can be derived from a trained model to automatically detect abnormal anatomies, meaning no specific classification model needs to be trained, and therefore annotated abnormal data are not required to build the framework. We proved that the stated methods provide a measurement of the abnormality of the model's output which is linked with the presence of malformations. We examined the approach with good results, not only on artificially generated data, but also in a large dataset of real clinical CT scans of patients with diverse inner ear malformations.

References

1. Baur, C., Graf, R., Wiestler, B., Albarqouni, S., Navab, N.: SteGANomaly: inhibiting CycleGAN steganography for unsupervised anomaly detection in brain MRI. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 718–727. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_69
2. Baur, C., Wiestler, B., Muehlau, M., Zimmer, C., Navab, N., Albarqouni, S.: Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain MRI. Radiol. Artif. Intell. **3**(3), e190169 (2021). https://doi.org/10.1148/ryai.2021190169
3. Amor, L. B., Lahyani, I., Jmaiel, M.: PCA-based multivariate anomaly detection in mobile healthcare applications. In: Proceedings of the International Symposium on Distributed Simulation and Real Time Applications (DS-RT), pp. 1–8 (2017). https://doi.org/10.1109/DISTRA.2017.8167682
4. Cairo/EG, R.Z.: Congenital inner ear abnormalities:a practical review. EPOS ECR 2019 / C-1911. https://doi.org/10.26044/ecr2019/C-1911, https://dx.doi.org/10.26044/ecr2019/C-1911
5. Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: a survey (2019), http://arxiv.org/abs/1901.03407
6. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995). https://doi.org/10.1006/cviu.1995.1004
7. Dhanasingh, A., et al.: A novel method of identifying inner ear malformation types by pattern recognition in the mid modiolar section. Sci. Rep. **11**(1), 1–9 (2021). https://doi.org/10.1038/s41598-021-00330-6
8. Gill, R.S., et al.: Deep convolutional networks for automated detection of epileptogenic brain malformations. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 490–497. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_56
9. Gower, J.C.: Generalized procrustes analysis. Psychometrika **40**(1), 33–51 (1975). https://doi.org/10.1007/bf02291478
10. Krenn, V.A., Fornai, C., Webb, N.M., Woodert, M.A., Prosch, H., Haeusler, M.: The morphological consequences of segmentation anomalies in the human sacrum. Am. J. Bio. Anthropol. **177**(14), 690–707 (2021). https://doi.org/10.1002/ajpa.24466, https://onlinelibrary.wiley.com/doi/10.1002/ajpa.24466
11. Leroy, G., Rueckert, D., Alansary, A.: Communicative reinforcement learning agents for landmark detection in brain images. In: MLCN/RNO-AI -2020. LNCS, vol. 12449, pp. 177–186. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66843-3_18

12. Diez, P. L., et al.: Deep reinforcement learning for detection of abnormal anatomies. In: Proceedings of the Northern Lights Deep Learning Workshop, vol. 3 (2022). https://doi.org/10.7557/18.6280
13. López Diez, P., Sundgaard, J.V., Patou, F., Margeta, J., Paulsen, R.R.: Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. In: MICCAI 2021. LNCS, vol. 12904, pp. 519–528. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_50
14. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
15. Seeböck, P., et al.: Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. IEEE Trans. Med. Imaging **39**(1), 87–98 (2020). https://doi.org/10.1109/TMI.2019.2919951
16. Sennarolu, L., Bajin, M.D.: Classification and current management of inner ear malformations. Balkan Med. J. **34** (2017). https://doi.org/10.4274/balkanmedj.2017.0367
17. Shin, H.-C., et al.: Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: Gooya, Ali, Goksel, Orcun, Oguz, Ipek, Burgos, Ninon (eds.) SASHIMI 2018. LNCS, vol. 11037, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00536-8_1
18. Trier, P., Noe, K. O., Sørensen, M.S., Mosegaard, J.: The visible ear surgery simulator, vol. 132 (2008)
19. Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B.: Multiple landmark detection using multi-agent reinforcement learning. In: shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 262–270. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_29
20. Yushkevich, P.A., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006). https://doi.org/10.1016/j.neuroimage.2006.01.015

PAPER E

Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images

Authors Paula López Diez, Josefine Vilsbøll Sundgaard, Jan Margeta, Khassan Diab, François Patou, and Rasmus R. Paulsen.

Journal Computerized Medical Imaging and Graphics. Volume 113, April 2024, 102343

Year 2024

Status Published

DOI https://doi.org/10.1016/j.compmedimag.2024.102343

Computerized Medical Imaging and Graphics 113 (2024) 102343

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Deep reinforcement learning and convolutional autoencoders for anomaly detection of congenital inner ear malformations in clinical CT images

Paula López Diez [a,*], Josefine Vilsbøll Sundgaard [a,f], Jan Margeta [c,e], Khassan Diab [d], François Patou [b], Rasmus R. Paulsen [a]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark
[b] Oticon Medical, Research & Technology group, Smørum, Denmark
[c] KardioMe, Research & Development, Nova Dubnica, Slovakia
[d] Tashkent International Clinic, Tashkent, Uzbekistan
[e] Oticon Medical, Research & Technology, Vallauris, France
[f] Novo Nordisk A/S, Denmark

ARTICLE INFO

Keywords:
Computed tomography
Anomaly detection
Deep reinforcement learning
Congenital malformation
Cochlear implant

ABSTRACT

Detection of abnormalities within the inner ear is a challenging task even for experienced clinicians. In this study, we propose an automated method for automatic abnormality detection to provide support for the diagnosis and clinical management of various otological disorders. We propose a framework for inner ear abnormality detection based on deep reinforcement learning for landmark detection which is trained uniquely in normative data. In our approach, we derive two abnormality measurements: D_{image} and U_{image}. The first measurement, D_{image}, is based on the variability of the predicted configuration of a well-defined set of landmarks in a subspace formed by the point distribution model of the location of those landmarks in normative data. We create this subspace using Procrustes shape alignment and Principal Component Analysis projection. The second measurement, U_{image}, represents the degree of hesitation of the agents when approaching the final location of the landmarks and is based on the distribution of the predicted Q-values of the model for the last ten states. Finally, we unify these measurements in a combined anomaly measurement called C_{image}. We compare our method's performance with a 3D convolutional autoencoder technique for abnormality detection using the patch-based mean squared error between the original and the generated image as a basis for classifying abnormal versus normal anatomies. We compare both approaches and show that our method, based on deep reinforcement learning, shows better detection performance for abnormal anatomies on both an artificial and a real clinical CT dataset of various inner ear malformations with an increase of 11.2% of the area under the ROC curve. Our method also shows more robustness against the heterogeneous quality of the images in our dataset.

1. Introduction

Inner ear malformation has been reported with an incidence of 20%–30% among children with congenital hearing loss (Brotto et al., 2021). The prevalence of bilateral congenital hearing loss is estimated as 1.33 per 1000 live births in North America and Europe, while in sub-Saharan Africa, the estimate is 19 per 1,000 newborns, and in South Asia up to 24 per 1,000 (Korver et al., 2017). Sensorineural hearing loss is generally detected early in countries with good access to healthcare services, which allows the prescription of interventions that mitigate the risk of abnormal social, emotional, and communicative development. These interventions include cochlear implant (CI) therapy, which is prescribed each year to about 80,000 infants and toddlers worldwide (Paludetti et al., 2012).

Radiological examination of children born with sensorineural hearing loss is key for an early diagnosis of congenital inner ear malformation. When the patient is treated with cochlear implant therapy, during the radiological examination the anatomy of the patient is evaluated to plan the surgical strategy. This surgery usually consists of drilling a precise tunnel from the surface of the scalp to the scala tympani in the cochlea where the implant is placed. The final location of the implanted electrode is critical for the patient's outcome (Chakravorti et al., 2019). Cases presenting congenital inner ear malformations raise

* Corresponding author.
E-mail address: plodi@dtu.dk (P. López Diez).

https://doi.org/10.1016/j.compmedimag.2024.102343
Received 21 September 2023; Received in revised form 25 January 2024; Accepted 28 January 2024
Available online 2 February 2024

many challenges during the planning and execution of CI surgery, often necessitating the surgeon to discover and adapt as the procedure progresses. Detecting and identifying such malformations from standard imaging modalities is a complex task even for expert clinicians. The detection and classification of the type of malformation is not a trivial task given the complexity of the anatomy and the great anatomical variation among malformations. Studies have defined several categories for these malformations, such as Sennaroğlu and Bajin (2017) which is one of the most popular works used for classifying congenital malformations of the inner ear. Dhanasingh et al. (2021, 2022) defined possible strategies for clinicians to detect congenital inner ear malformations based on the visual exploration of CT scans involving explicit measurements and humans' natural ability for pattern recognition. The strategy described in Dhanasingh et al. (2022) is based on visualizing the cochlea in two specific planes (oblique-coronal and mid-modiolar) and following three steps: the cochlear A and B distances defined as Escudé et al. (2006), the number of cochlear turns, and a visual analysis based on an assessment of resemblance to different objects such as the Aladdin's lamp and a side view of a dog's face. This methodology provides empirical guidelines for clinicians to detect these malformations based on hand-crafted features. As with any human-based image interpretation method, it is time-consuming and can be subject to the clinician's subjectivity during evaluation.

In López Diez et al. (2022b), we introduced the first automated approach for detecting inner ear congenital malformations. In that approach, we used a deep reinforcement learning (DRL) model trained for landmark localization exclusively in normal anatomies to derive two anomaly measurements: the first was based on the variability of the predicted configuration of the landmarks in a subspace formed by the point distribution model of the normative landmarks' location using Procrustes shape alignment and principal component analysis (PCA) projection. The second measurement was based on the distribution of the predicted Q-values of the model for the last ten states before the landmarks were localized. In the present paper, we build on our prior work and compare this approach to a 3D convolutional autoencoder approach in which the patch-based reconstruction error is used for anomaly detection in 3D images, as described by Sato et al. (2018).

This journal paper presents a significant extension of the conference work presented at MICCAI 2022 (López Diez et al., 2022b). We have extended the literature review, especially for unsupervised anomaly detection (UAD) in medical images. Furthermore, we have chosen to use the MARL architecture and not the communicative version (C-MARL), as it has proven best for this task based on our findings, and we have closely benchmarked our method against another state-of-the-art 3D-based approach for UAD in both artificially generated anomalies and clinical images of congenital inner ear malformations. This helps us to understand how current approaches tested on brain datasets might perform in different and more challenging datasets, as is the case with our rare clinical dataset. Furthermore, our approach allows us to assess how the performance generalizes from artificially generated datasets and heterogeneous real clinical scans to scans from a very specific and controlled environment.

2. Background

Recently, machine learning has enabled the development of automated medical image analysis methods that achieve great levels of performance in the detection of abnormal anatomies and other types of anatomical anomalies (Sajid et al., 2019; Wang et al., 2022). Despite their outstanding performance, these methods, which are based on supervised learning, have a major disadvantage: they require large labeled datasets that faithfully represent the spectrum of possible anomalies. These datasets are scarce, and costly to obtain, especially for rare diseases such as congenital inner ear malformation. Furthermore, it is very difficult to predict how these supervised models will behave with new unseen data. Lately, some deep learning approaches that seek to enhance UAD have been introduced. These new deep-learning-based UAD methods resemble the clinical approach to image exploration and can detect anomalies without prior knowledge about the anomalies' appearance.

Historically, UAD was based on statistical models (Van Leemput et al., 2001), out-of-the-distribution techniques (Prastawa et al., 2004; Allenby et al., 2021), hand-crafted features (Martins et al., 2020), content-based retrieval or clustering (Taboada-Crispi et al., 2009). Nowadays, approaches based on isolation forest (Liu et al., 2012), like (Hariri et al., 2021; Xu et al., 2023a) for UAD, have gained significant popularity and demonstrate high performance. However, these approaches rely on the anomalies being distinct and sparse, being this last one a condition not met by the datasets utilized in this study. Furthermore, these approaches have been used only in features or lower-dimensional spaces derived from CT images as in the work presented by Hainan et al. (2019) and Welch et al. (2020). Nevertheless, they are not yet suitable for direct application to high-dimensional clinical data, such as CT or MRI images. This misalignment contradicts our goal of utilizing an approach directly applicable to the entire CT image. New deep learning techniques are being tested for UAD as the ones presented in Wang et al. (2023), Xu et al. (2023b). However, in the medical image domain, most of the UAD methods used for 3D images are based on an autoencoder approach. The underlying concept is to learn an implicit and synthesized representation of a certain type of image in normative samples and use the difference between the original image and the one produced by the generative branch of the model to estimate the probability of the given sample being an anomaly. Different versions of convolutional autoencoders (CAE) (Baur et al., 2019; Sato et al., 2018; Atlason et al., 2019; Astaraki et al., 2022) and variational autoencoders (VAE) (Chen et al., 2020; Pawlowski et al., 2018; Zimmerer et al., 2019; Astaraki et al., 2022) have been tested for UAD in medical images. These are popular strategies to tackle unsupervised anomaly segmentation by modeling the distribution of normal images. In a similar manner, different GANs-based approaches (Schlegl et al., 2019; Baur et al., 2020; Sun et al., 2020; Schlegl et al., 2017) have been used for this purpose as well. More recently, autoencoders with transformers (Pinaya et al., 2022b) and diffusion models (Pinaya et al., 2022a; Wolleb et al., 2022) have been proposed for UAD. Finally, attention-map-based approaches such as the ones presented by Silva-Rodríguez et al. (2022) and Venkataramanan et al. (2020) are also being used for UAD on medical images.

Besides the work by Sato et al. (2018) and Pinaya et al. (2022b), all the previously mentioned works have proposed 2D-based approaches, even though some are used for processing volumetric data, mainly MRI and CT scans. These 2D approaches do not exploit all the implicit information of the 3D scans, even if they are computationally more efficient. This inability to exploit all the information is problematic in UAD for complex anatomies, such as the inner ear, for which 3D spatial information is essential in correctly analyzing the internal structures, which are small and interconnected with a high degree of curvature. Despite their success, transformer-based approaches, such as Pinaya et al. (2022b), still have some weaknesses intrinsic to their autoregressive nature, as it is the fixed order of sequence elements that creates a bias to attention. This problem is more noticeable in 3D images, where even more transformers might be required to achieve good coverage of the image context given the images' higher dimensionality. Therefore, we chose to compare our proposed 3D-based UAD method with the asymmetric 3D convolutional autoencoder described by Sato et al. (2018) as we consider it the best-suited approach for direct comparison with a low demand of computational resources.

In a similar fashion, we tackle the anomaly detection problem with a parametric approach instead of a classification one, in an attempt to build implementations that move toward more interpretable results. We decided to use the landmark-based approach as an object search problem using a DRL approach trained in normative data to derive implicit information that can be used for UAD. The implicit information

used is of two different types. The first type is based on the variability of the predicted configuration of the pre-specified landmarks in a subspace defined by the point distribution model of the normative location of the landmarks using Procrustes shape alignment and Principal Component Analysis projection. The second type is based on the distribution of the predicted Q-values of the model for the last ten states before the landmarks are localized as an agent hesitation measurement. Landmarks located with DRL have been used for anomaly detection by Bekkouch et al. (2022), where a method for abnormality detection in 2D X-ray images of the hip is proposed. This method is however not a UAD as it is not unsupervised because the agents are trained to localize the landmarks in abnormal cases and their prediction is then used to estimate if they fall within the healthy population expected inter-landmark relationship. This approach is therefore limited, as are all the supervised methods, by the size and representativity of the dataset used in comparison to all the possible abnormal cases.

Deep reinforcement learning consists of the use of deep learning to solve a reinforcement learning problem. Deep reinforcement learning has been applied to medical images with great success over the last years for parametric medical image analysis, optimization problems, and image classification (Zhou et al., 2021). Even though the automatic detection of the different types of inner ear malformation is, by nature, a classification problem, due to the lack of availability of representative and heterogeneous datasets that faithfully represent the full spectrum of these congenital malformations, we use a parametric approach (landmark detection) in normative data to derive implicit information that can potentially detect an anomaly in the anatomy in an unsupervised manner.

Many existing UAD approaches focus on the detection of brain cancer (Nazir et al., 2021), where cancer can appear at almost random locations in the entire brain. The goal of those approaches is to detect where the brain looks abnormal compared to a normative population. In our case, we are specifically looking at a very small anatomical region, the cochlea, that when abnormal might have a very different overall appearance compared to a normative population. Our assumption is that the configuration of a limited number of anatomical landmarks can provide the necessary information for anomaly detection. For a brain scan with a randomly placed tumor, a landmark-based UAD would require an extensive amount of anatomical landmarks and we do not believe that our DRL approach would be suitable for that task.

3. Materials and methods

3.1. Data

In this study, two different datasets have been used to evaluate the different methods: an artificial dataset and a clinical dataset.

The artificial dataset consists of 119 clinical CT scans of patients with normal inner ear anatomy. The cochlear structure presents some variability in normative patients but it is fairly consistent across this population (Demarcy et al., 2017). This dataset is composed of images from diverse CT scanners which were cropped to a standard view and orientation of the region of interest (ROI) of 32.1^3 mm^3 using the Nautilus software (Margeta et al., 2022) and their proposed orientation. These images were labeled with five anatomical points of interest for nerve characterization (López Diez et al., 2021), an example of which is shown in Fig. 1. To test our approach, we synthetically generated abnormal inner ear CT scans from the original images by removing the cochlea (simulating cochlear aplasia), thus generating corresponding pairs of normal and abnormal CT scans with the same surrounding structures. The cochlea was segmented using ITK-SNAP software (Yushkevich et al., 2006) and then replaced by Gaussian noise with mean and standard deviation estimated from the intensities of the tissue surrounding the segmentation (López Diez et al., 2022a).

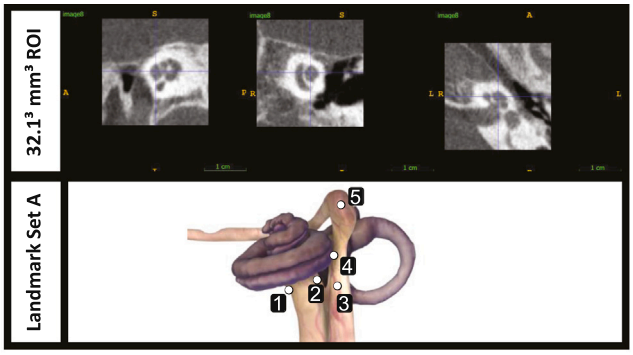


Fig. 1. Artificial dataset. Top: Example of CT scan from the artificial dataset with a ROI of 32.1^3 mm^3 and isotropic spacing of 0.2 mm. Bottom: Landmark set A: **1, 2** - Opposite sides of bony cochlear nerve canal in axial view. **3** - Facial Nerve (FN) exiting the internal acoustic canal. **4** - Closest point of FN and cochlea. **5** - Geniculate ganglion of the FN.
Source: Figure edited from Trier et al. (2008).

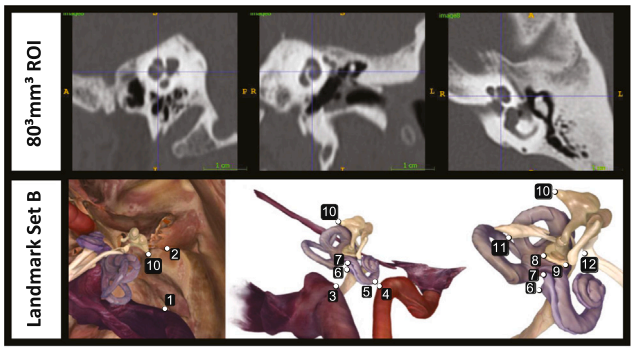


Fig. 2. Clinical dataset. Top: Example of CT image from the clinical dataset with a ROI of 80^3 mm^3 and isotropic spacing of 0.15 mm. Bottom: Landmark set B: **1** - Sigmoid sinus (closest point to the external acoustic canal). **2** - External acoustic canal (closest point to).sigmoid sinus **3** - Jugular bulb (closest point to round window). **4** - Carotid artery (closest point to basal turn of the cochlea). **5** - Basal turn (closest point to JB). **6,7** - Anterior and posterior edges of RW. **8,9** - Anterior and posterior crus of staples. **10** - Short process of incus. **11** - Pyramidal process **12**- Cochleariform process.
Source: Figure edited from Trier et al. (2008).

Our second dataset, the clinical dataset, consists of 300 anatomically normal CT scans from heterogeneous sources and 122 CT scans of inner ears that present different types of inner ear malformations. The ROI extraction for this dataset was done using the methodology described by Radutoiu et al. (2022) using anatomical points of interest that were not involved in the anatomy of interest in order to allow for a standardized and robust image orientation regardless of the appearance of the inner ear region. A greater ROI of 80^3 mm^3 was selected for this dataset in order to contain all the anatomical points of interest for CI therapy. For this dataset, as shown in Fig. 2, twelve anatomically relevant landmarks were carefully designed and annotated in a randomly selected subset of 160 CT scans of anatomically normal cases in collaboration with our clinical partner, an ENT surgeon specialized in CI therapy in abnormal anatomies.

3.2. DRL for landmark localization

Reinforcement learning is a computational approach for learning an optimal policy by interacting with the environment E. An agent observes its current state, s, and chooses an action, a, from its set of possible actions, A, and the environment returns a reward, r, which characterizes the quality of the action chosen. For landmark localization in 3D images, the problem is defined as the environment, E, being

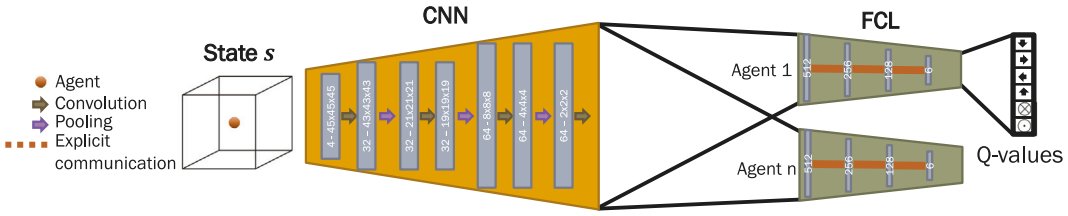


Fig. 3. Diagram of the multi-agent reinforcement learning architecture used. The input is a patch centered in the current agent's position (state). In yellow we show the convolutional neural network which extracts the relevant features of a certain patch. Those features are then passed to each corresponding agent which consists of a set of fully connected layers (in green) that map those features to the estimated expected reward (Q-values) of each of the possible actions (up, right, left, down, forward, or backward). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the 3D image and the agent is a physical location within the image. The state, s, is a patch of the image centered in the agent's location and the action set, A, is the movement in one of the six Cartesian directions (up, down, left, right, forward, and backward). The reward, r, is defined as the difference between the distance to the target landmark's location after the last action and in the previous state, meaning it is a positive value if the agent is moving closer and a negative one if it has moved away from the target landmark's location. The agent's goal is to learn an optimal policy that maximizes not only the immediate reward but also the subsequent future rewards.

The expected reward of taking a certain action given a state is defined as the Q-value. In deep reinforcement learning, this Q-value is estimated using a Deep-Q-Network which takes the current state as its input, and outputs the Q-value associated with each possible action. The architecture of the Deep-Q-Network used for landmark location resembles a typical image classification architecture. We use a MARL (Vlontzos et al., 2019) architecture which includes a set of fully connected layers for each agent. An illustration of the architecture used can be seen in Fig. 3. The agents all share the same convolutional layers (implicit communication), meaning the feature extractor is common to all the agents, but each agent has its own set of fully connected layers. While explicit communication was introduced in the C-MARL model proposed by Leroy et al. (2020), we do not include explicit communication, as we found in López Diez et al. (2022b) that it is not beneficial for anomaly detection.

We employ a multi-scale approach in which the artificial agent is trained not only to distinguish the target within the anatomy but also to learn and follow an optimal navigation path to the target landmark location in the 3D image, as introduced by Ghesu et al. (2019). The agent's search starts at the coarsest scale level with a global context and continues across three different scales, capturing increased levels of detail when transitioning to finer scales. This resembles the human approach to landmark detection in medical images, starting with a big field of view and localizing the region of interest where the clinician zooms in and continues looking for the specific features of that specific landmark.

3.2.1. PCA shape distance method

The landmark locations defined in Figs. 1 and 2 present a consistent spacial configuration among patients with a normal inner ear anatomy. The assumption is that when the anatomy does not resemble the anatomical appearance and configuration the agents have been trained with, the final location will deviate significantly.

To evaluate the configuration of the predicted landmark locations, we use a point distribution model (PDM) following the approach presented by Cootes et al. (1995). A full set of landmark locations in an image is denoted as a shape with a point correspondence across all shapes in the training data, and this correspondence is known. Firstly, the alignment between all the shapes from normative data is derived using Procrustes analysis (Gower, 1975). By using this transformation, we obtain a PDM that represents shape variability within the ROI for normal anatomies and is invariant to size and orientation. We can thus derive the mean shape $\bar{x}$ from this model, followed by a PCA of the shape variation (Cootes et al., 1995).

From this analysis, we obtain the matrix, Φ, which is a set of the principal components describing the variability of the shape in the healthy dataset. Based on this, a new shape, x', can be defined as: $x' = \bar{x} + \Phi b$, where the vector b defines the weights controlling the modes of shape variation and Φ contains the first t principal components, which we defined as the lower possible t such that 90% of the shape variability is contained in the Φ matrix. For the artificial dataset, shown in Fig. 1, it was found that $t = 6$ was enough, while for the clinical dataset, shown in Fig. 2, $t = 11$ was found sufficient.

When a new set of landmark locations is predicted, the new shape, x', can be aligned to the mean shape, $\bar{x}$, and be approximated by the PDM model by projecting the residuals from the average shape into principal component space: $b = \Phi^T(x' - \bar{x})$.

The vector b describes the shape coordinates in the PCA space. In this space, we evaluate the distance between the different shapes predicted by the model. We then compute the Euclidean distance between the projected shapes as

$$d_{ji} = \|b_i - b_j\|_2 \tag{1}$$

which quantifies the variation of all the different shapes predicted for a certain image, where i and j represent two different shapes within the same image. Finally, we compute the standard deviation of this distribution of distance values for a certain image in the following manner

$$D_{\text{image}} = \sqrt{\frac{\sum |d - \bar{d}|^2}{n}} \tag{2}$$

where n is the number of different distances within one images. D_{image} measures the level of agreement among the multiple predictions computed in the PCA space defined by normative shapes. A sketch of this approach is shown in Fig. 4(I).

3.2.2. Q-value history distribution method

It is our assumption that the expected rewards (Q-values) predicted during the landmark location search process, in which the agent is navigating the image, could represent the degree of confidence (or, defined anthropomorphically, of hesitation) an agent has about the final landmark location. This hesitation measurement should be highly correlated with the anatomical appearance, meaning it could be used to detect anomalies in the anatomy where the landmarks are localized.

Given a certain state of the agent resembles the normal anatomical configuration of such a region, we expect that the Q-values will present a uniform distribution as the agent should not expect a high reward for moving in a certain direction. On the other hand, when the anatomy of the current state does not resemble what the agent is looking for, the Q-values should be less uniformly distributed, pushing the agent to move away from the current location. We define a measurement of the variability within the distribution of the predicted Q-values of the action set in the last stages prior to the final landmark location. To compute this hesitation or uncertainty measurement, we collect the buffer of predicted Q-values of the last 10 states of the agent, which

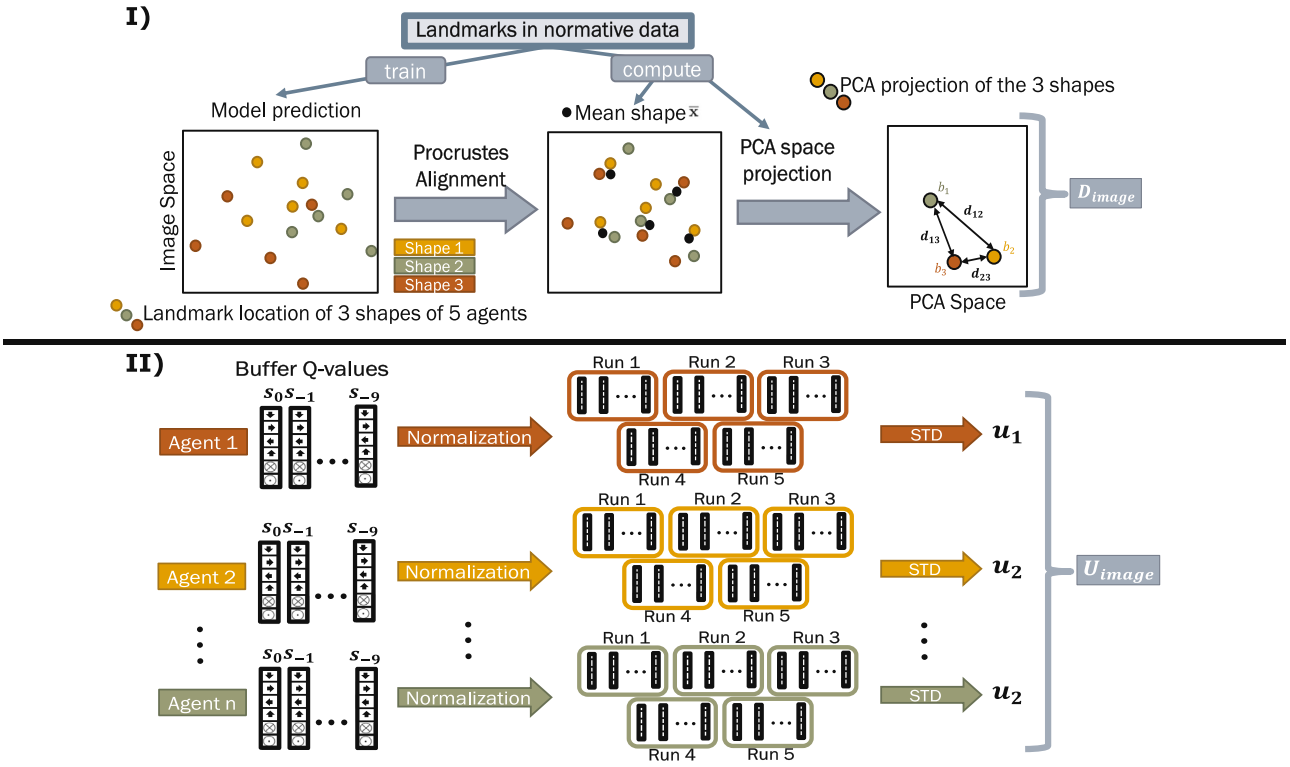


Fig. 4. (I) Diagram of the PCA shape distance method and how D_{image} is computed. (II) Diagram of the Q-value history distribution method and how U_{image} is computed.

have empirically been found sufficient to define the later stages of the landmark search procedure. The normalization of the Q-values is done using the last 10 states of an agent, these values are divided by the biggest Q-value of that agent in that run and the standard deviation is computed using those normalized values over all the runs, which is the uncertainty measurement of that landmark, u_n. These uncertainty measurements are then joined together by computing the norm of the vector containing the u_n of all the landmarks in that image into a single value per image as follows

$$U_{\text{image}} = \sqrt{\sum_n u_n^2} \tag{3}$$

We check the uniformity of the Q-values distribution for each landmark independently and not over all the landmarks in one image because different landmarks have specific anatomical relevance. An overview of this method is shown in Fig. 4(II).

3.2.3. Combined anomaly measure

To evaluate whether the two proposed methods could complement each other, the joint performance is also taken into account. Due to the magnitude of the measurements being different from one another, we introduce a weighting factor to obtain a more representative combination of both methods. This weighting factor is computed by estimating the median of both D_{image} and U_{image} over all the training images in order to determine the intrinsic magnitude difference between both measurements. The weighting factor is then defined as

$$w = \frac{\text{median}(D_{\text{training}})}{\text{median}(U_{\text{training}})} \tag{4}$$

The combined measure, which can be used to analyze the joint performance, is therefore defined as:

$$C_{\text{image}} = \sqrt{D_{\text{image}}^2 + (wU_{\text{image}})^2} \tag{5}$$

to analyze the joint performance.

3.3. 3D convolutional autoencoder

An autoencoder is an unsupervised learning algorithm that learns an identity mapping of the input by minimizing the loss function between the input and its reconstructed output. It is based on both an encoding and a decoding phase. In the encoding phase the original image, $I \in \mathcal{R}^D$, is compressed into a feature vector, $y \in \mathcal{R}^d$, that can be reconstructed back to the original space $\hat{I} \sim I$, given $D \gg d$ in the decoding phase. Autoencoders are very well suited for different tasks such as anomaly detection but also for simplifying the process of feature engineering in machine learning studies, as well as for dimensionality reduction, denoising data, generative modeling, and even pretraining deep learning neural networks (Lopez Pinaya et al., 2020).

Convolutional autoencoders (CAEs) (Masci et al., 2011) are based on the same principles but use deep convolutional layers to perform the dimensionality reduction. The local connectivity of convolutional layers enables the CAE to extract local and hierarchical features capturing the global feature of the input by combining the local features. These local connections require less computational cost than full connections. Pooling layers are used to reduce the input size and to add robustness to shift and position variance. 3D-CAE is an extended CAE composed of 3D convolution and pooling layers, applicable to volumetric data (Arai et al., 2018). We use the asymmetric architecture proposed by Sato et al. (2018), which they employed for anomaly detection in emergency head CT volumes. The architecture consists of a contracting path (3D-CNN) and a reconstructive path (3D-deCNN). Details about the architecture are shown in Fig. 5.

We consider the reconstruction error as the squared difference in intensity between input and output $\mathcal{E} = (I - \hat{I})^2$. The Mean Squared Error (MSE) is used as the loss function to train the network.

3.3.1. Abnormality measurement

Given that the CAE has only been trained on anatomically normal images, it is assumed that the model will learn how to efficiently synthesize such images. This implies that some possible implicit patterns

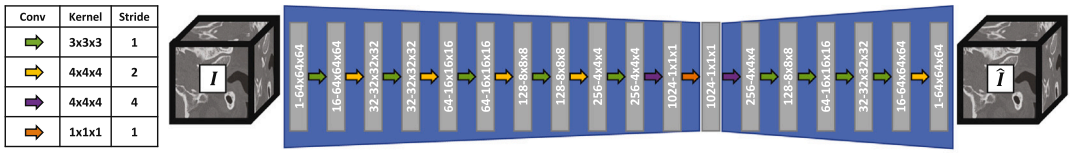


Fig. 5. Architecture of the asymmetrical 3D convolutional autoencoder used for anomaly detection where the different convolutional layers are characterized. I is the input image which is a CT scan of the inner ear and $\hat{I}$ is its generated version after the image has been synthesized into the feature vector of size 1024.

that are consistent throughout the whole dataset might not be a specific part of the encoding, given that they are always present in a similar way. This is the case for the cochlear structure, whose anatomical structure is complex yet very consistent across normative subjects. A higher reconstruction error is expected for images that are anatomically different to the ones in the training set. The chosen error measurement for the CAE model is the patch-based MSE, as used in the study by Sato et al. (2018). The abnormality measurement is defined as the MSE of a patch and all the values for the patches of a certain image are concatenated in a vector denoted a_{image}. The abnormality measurement of a certain case is thus defined as the maximum abnormality measurements over all the patches in an image, defined as

$$A_{\text{image}} = \max\{a_{\text{image}}\} \tag{6}$$

3.4. Experiments

The 119 anatomically normal cases of the artificial dataset were randomly split into: 87 cases for training, 10 for validation, and 22 for testing. The test set comprises those 22 cases and their corresponding artificial anomalies generated with the transformation explained in Section 3.1, resulting in 44 images for evaluation. The clinical dataset was randomly split into a training set of 160 (anatomically normal) images, a validation set of 18 (anatomically normal) images, and 244 images for testing (122 anatomically normal and 122 with congenital malformations). Both the MARL models for landmark localization and the 3D-CAE model were trained using the same split of the data. The best-performing model on the validation set was chosen as the final model for evaluation.

All the models were trained end-to-end on a Titan X 12 GB GPU. The MARL models were trained with one agent per landmark, meaning five agents for the artificial dataset and 12 for the clinical dataset. The models were evaluated over five inferences in order to compute the corresponding anomaly metrics introduced in Eqs. (2), (3) and (5). The approximate training time for the two MARL models used was five days. In the training process, the final state is reached when the distance to the landmark is ≤ 1 voxel using the ϵ-greedy search strategy (Watkins, 1989). During inference, the agent's oscillation is used to finalize the search. The forgetting factor γ is set to 0.9 as this has been empirically found to be the best-performing option. We used a multi-scale approach with three isotropic resolutions: 0.5, 0.250, and finally 0.125 mm and a frame history of 4 observations.

The 3D-CAE model software was developed using PyTorch (Paszke et al., 2019) building on top of the MONAI software (MONAI-Consortium, 2022). With a batch size of 8 and a patch size of $128 \times 128 \times 128$ voxels. We used a learning rate of 5.10−5 and AdamW optimizer for training.

4. Results

Processing results for the artificial dataset are shown in Fig. 6 and the ones for the clinical dataset in Fig. 7. In Figs. 6(a) and 7(a) we have included the Receiver Operating Characteristic (ROC) curve which represents the trade-off between true positive rate (TPR), also known as sensitivity, and the false positive rate, which is equivalent to 1-specificity, for all the possible thresholds of the binary classification. We also include the maximum accuracy metric and the Area under the Curve (AuC) as an overall summary of anomaly detection performance that can be observed in Table 1.

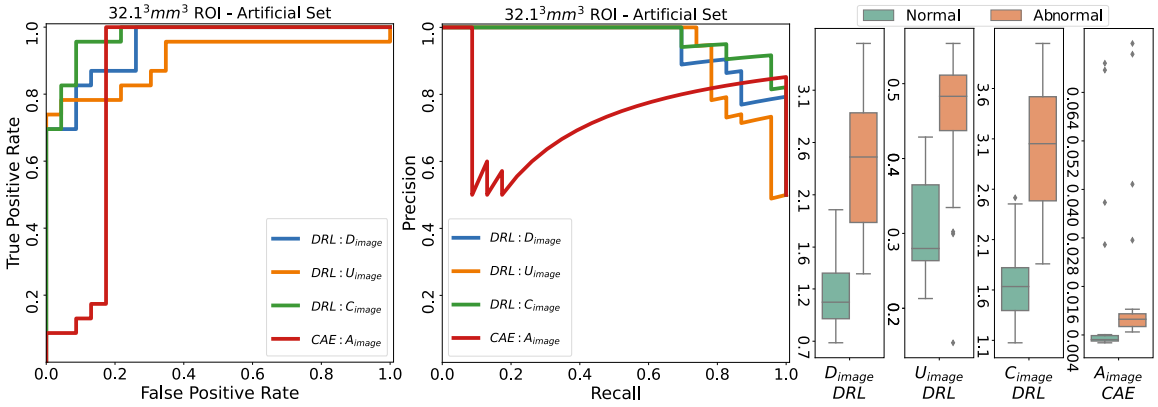
In Figs. 6(b) and 7(b) the precision–recall curve of the anomaly detection is represented. Even though we have created a perfectly balanced test set for both experiments, it is relevant to get a better overview of the classifier performance. These curves represent the relationship between recall (TPR) and precision which measures the fraction of examples classified as anomalies that are truly anomalies. We have also computed the maximum f1-score for each of the methods which is shown in Table 1.

Finally, in Figs. 6(c) and 7(c), boxplots of the distribution of the different anomaly measurements of each method are represented for the artificial and the clinical dataset respectively.

Fig. 8 shows an original image, I_{original}, from the test set and its corresponding reconstructed version, $\hat{I}_{\text{original}}$, followed by its artificially generated abnormal version, $I_{\text{artificial}}$, and its reconstruction, $\hat{I}_{\text{artificial}}$. Furthermore, we also display the reconstruction error, $\mathcal{E}$, for both the original, $\mathcal{E}_{original}$, and the artificial image, $\mathcal{E}_{artificial}$. It can be observed in the reconstructed artificial image, $\hat{I}_{\text{artificial}}$, that even though the cochlea had been artificially removed from the input image, $I_{\text{artificial}}$, it generates a normal cochlear shape, very similar to the one shown in the original image from the corresponding pair, I_{original}. This shows that the model has indeed learned the implicit representation of the normal cochlea and always reconstructs an image with normal anatomy. The resemblance between both outputs is very clear and we can understand the artificial image will have a higher reconstruction error, especially in the cochlear region which could be used to segment or indicate the region of the image that presents the anomaly.

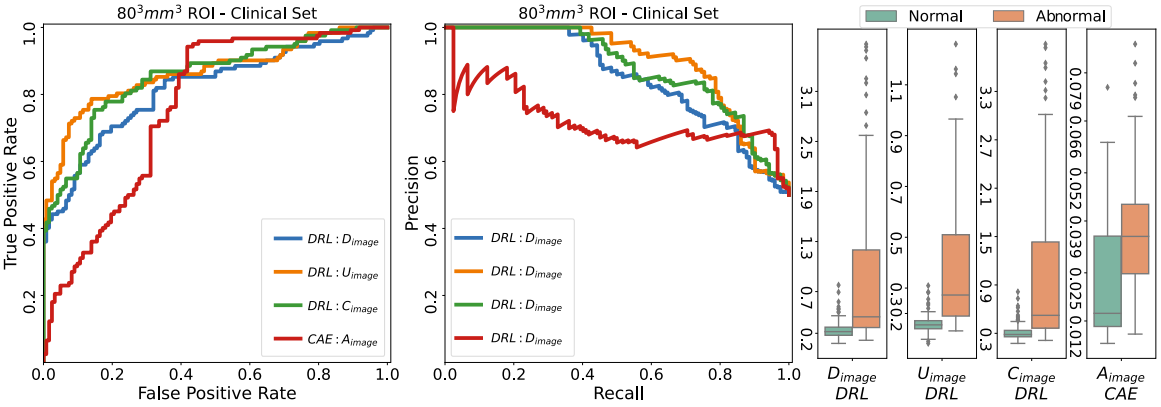
However, is it important to note that the reconstructed images in Fig. 8 present a more smooth overall appearance with less noise than the input images. As previously mentioned, we collected both of our datasets from different clinics meaning the images come from different CT scanners and present different image quality levels, thus the datasets are quite heterogeneous. When visually analyzing the results in the artificial dataset, one notes that the more noisy scans present a greater and generalized reconstruction error due to the denoising effects of the autoencoder. This behavior is the main reason behind the results shown in Fig. 6, where a better result was expected for the 3D autoencoder given the smaller ROI and the artificially introduced anomaly that is more extreme than most of the real clinical cases. In addition, the fact that the artificial dataset presents corresponding pairs of scans with and without the malformation for evaluation allows for an analysis of the pair-wise behavior, where we also observe that the A_{image} measurement is always greater in the corrupted image but a global threshold is not successful for classification. This can be observed in Fig. 6(c) where the corresponding outliers that present a greater reconstruction error in the normal anatomies have their corresponding pairs for the abnormal cases which have a slightly greater value of A_{image}. Therefore setting a general threshold for binary classification is quite challenging for a heterogeneous dataset, but very feasible for a model implemented exclusively for a homogeneous dataset for a specific CT scanner. In Figs. 6(a) and 6(b) we can see that the CAE-based model only detects 10% of the anomalies if any false positive is tolerated.

Our DRL-based methods generally outperform the 3D-CAE approach, which is observed in the better performance curves in both Figs. 6 and 7. However, for both experiments, we see that there is a tendency



(a) Receiver Operating Characteristic (ROC) curve. (b) Precision-Recall curve. (c) Distribution of anomaly measurements.

Fig. 6. Evaluation in the artificial dataset. Performance metrics obtained with each of the methods based in DRL (green, blue and orange) against the performance of the 3D-CAE method (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Receiver Operating Characteristic (ROC) curve. (b) Precision-Recall curve. (c) Distribution of anomaly measurements.

Fig. 7. Evaluation in the clinical dataset. Performance metrics obtained with each of the methods based in DRL (green, blue and orange) against the performance of the 3D-CAE method (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

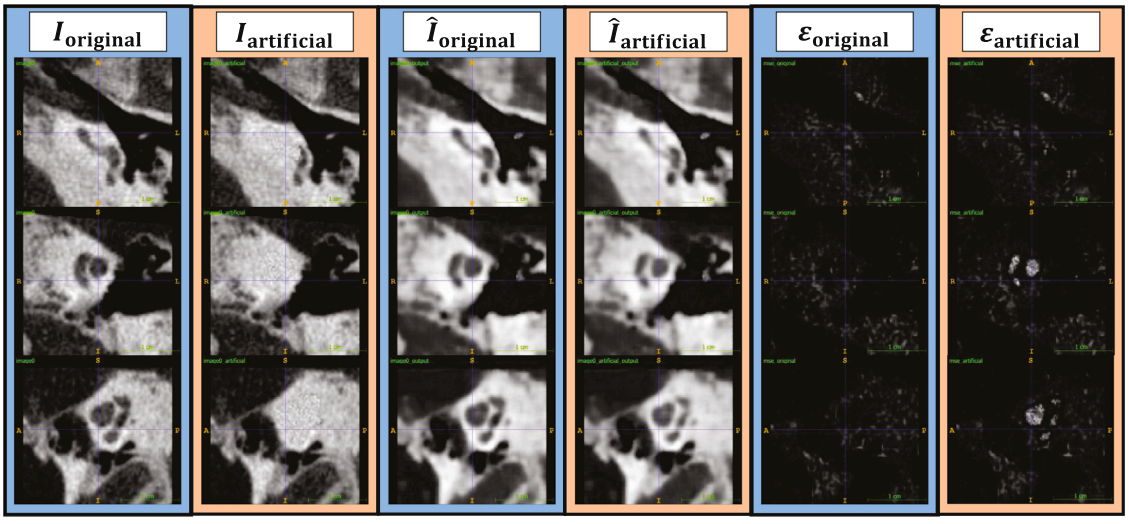


Fig. 8. 3D-CAE test examples from the artificial dataset. I_{original} is the input CT clinical image and $I_{\text{artificial}}$ is its corresponding corrupted version (in-painted cochlear structure). $\hat{I}_{\text{original}}$ and $\hat{I}_{\text{artificial}}$ are the corresponding reconstructed version of the input images. $\mathcal{E}_{\text{original}}$ and $\mathcal{E}_{\text{artificial}}$ present the reconstruction error between input and output for each case.

that if false positives are not tolerated, then a DRL approach clearly outperforms the 3D-CAE, but if a high rate of normal cases detected as anomalies is tolerated (around 50% for the clinical dataset as seen in Fig. 7(a)), then the 3D-CAE has a similar or slightly better accuracy depending on the threshold and dataset. Given this, it is natural that in Fig. 7(b) we see that the 3D-CAE approach suffers a smaller drop in precision as the recall is increased, but an overall significantly lower precision for recall values from 0 to approximately 0.8.

When comparing the results in Figs. 6 and 7, a similar trend is observed for both experiments. This proves that the hypothesis tested

Table 1
Summary of evaluation metrics for the different methods for anomaly detection in the inner ear anatomy in artificial and clinical datasets.

	Artificial dataset			Clinical dataset		
	AuC	Max. f1-score	Max. acc.	AuC	Max. f1-score	Max. acc.
DRL: D_{image}	0.95	0.88	0.87	0.82	0.77	0.76
DRL: U_{image}	0.90	0.86	0.87	0.87	0.82	0.82
DRL: C_{image}	0.97	0.94	0.93	0.86	0.80	0.80
3D-CAE: A_{image}	0.85	0.92	0.91	0.76	0.80	0.76

with a sparser dataset that contains a more reduced field of view and where the images have been artificially transformed into the more severe type of malformation generalizes well to a complicated clinical dataset with different types of malformations with a greater field of view of the image.

For our DRL approach, we observe that the combined anomaly measure, C_{image}, shows improved performance on the artificial dataset as shown in Fig. 6, and a very close performance to the best-performing method on the clinical dataset as seen in Fig. 7. These results are also shown in Table 1. We consider that the combination, C_{image}, should be used as a more reliable measurement, even though in the clinical dataset we observe a small better performance from the U_{image} measurement.

It is also interesting to note how the different evaluation metrics shown in Table 1 vary for the different methods and datasets. There is a clear drop in performance between the artificial and the clinical dataset, as was expected, given the increased complexity of the task. If we look at the AuC, the difference in performance between any of the DRL methods and the 3D-CAE is very clear ($11,2\%$ improvement on average). Meanwhile, we see similar values for the maximum f1-score and accuracy, which sometimes is higher for the 3D-CAE than for some of the DRL approaches, however, the C_{image} always presents similar or superior metrics than A_{image}.

5. Discussion

Our DRL-based method outperforms the 3D-CAE mostly due to a better adaptation to heterogeneous datasets which are typical in a clinical context. In our experiments, the autoencoder presents a bigger sensitivity toward the nature and origin of the image. The search agents are more robust to the quality difference between images, even though they are trained to choose an optimal action given a certain crop of the image, the appearance is not directly correlated with the loss function, as it is in the case of the autoencoder, nor is it directly linked with the final anomaly measurement. Of course, the image's quality also affects the DRL approach's performance, as it would do for a clinician who is searching for the location of a certain number of landmarks. Images of a lower quality are still more challenging for the DRL approach because the quality of the extracted features will be affected by this, but not to the same extent as the 3D-CAE approach, as can be observed especially when analyzing the performance in the artificial set where the 3D-CAE shows a lower tolerance towards noisy scans, as explained in Section 4.

Both approaches have the potential to be used as a more interpretable anomaly detector rather than a basic classifier because both approaches contain spatial information about the original image that can be exploited. In the case of the 3D-CAE model, the reconstruction error $\mathcal{E}$ can be seen as a map of the abnormal areas indicated by a higher error, which can be highlighted to the clinician as areas of interest. For the DRL approach, the use of specific landmarks that are key for the studied anatomy provides information on the relative points of interest for each case. In both abnormality measurements D_{image} and U_{image}, the information of each agent (corresponding to one landmark) could be used to indicate which region of the image contributes more to the final measurement. This potential for interpretability allows for highlighting regions of the image that have influenced the decision. Clinicians could look into this area of interest and detect something that might have otherwise been overseen, such as an anomaly or the reason for a falsely detected anomaly, which might be, for example, an artifact in the image.

For our DRL-based approach, the normative images used for training must be annotated with all the landmarks of interest, which can be time-consuming. The 3D-CAE approach does not require pixel-level annotations but only confirms that the image contains normal anatomy. However, the 3D-CAE approach does require a more strict standardization of the input image. The DRL approach is less sensitive to scale variations or small differences in orientation given its multiscale approach. This supports the previously introduced idea that the 3D-CAE approach will be better suited for homogeneous datasets, while the DRL-based approach generalizes better for more heterogeneous datasets.

6. Conclusion

We have shown that congenital inner ear malformations in CT images can be automatically detected by training a DRL model exclusively on normative data and evaluating the output variability of its implicit information. This information contains the relative position of the predicted landmarks' location over different runs/agents in a subspace defined by the normative annotations as well as the distribution of the Q-values of the last iterations of the agents as a measurement of the uncertainty of the final location. We also compare the proposed approaches with an asymmetric 3D-CAE, which is based on a 3D-approach for volumetric data. We compare the performance between both methods and analyze the results obtained not only on artificially generated data but also in a large dataset of real clinical CT scans of patients with diverse inner ear malformations from several different clinics. The DRL approach outperforms the 3D-CAE method in both datasets, mostly because it presents a higher tolerance towards heterogeneous real clinical scans from different sources. We believe that the presented DRL approach could be readily adapted to other anatomies prone to complex anatomical anomalies. This could include, but not be limited to, congenital heart disorders or complex spine compressions.

CRediT authorship contribution statement

Paula López Diez: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition. **Josefine Vilsbøll Sundgaard:** Methodology, Software, Writing – review & editing. **Jan Margeta:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **Khassan Diab:** Conceptualization, Resources, Supervision. **François Patou:** Funding acquisition, Project administration, Supervision, Writing – review & editing. **Rasmus R. Paulsen:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Paula Lopez Diez reports financial support was provided by William Demant Foundation.

Data availability

The authors do not have permission to share data.

Acknowledgments

We would like to thank William Demant Fonden (Denmark) for financially supporting this study.

References

Allenby, M.C., Liang, E.S., Harvey, J., Woodruff, M.A., Prior, M., Winter, C.D., Alonso-Caneiro, D., 2021. Detection of clustered anomalies in single-voxel morphometry as a rapid automated method for identifying intracranial aneurysms. Comput. Med. Imaging Graph. 89, 101888. http://dx.doi.org/10.1016/j.compmedimag.2021.101888, URL https://www.sciencedirect.com/science/article/pii/S0895611121000367.

Arai, H., Chayama, Y., Iyatomi, H., Oishi, K., 2018. Significant dimension reduction of 3D brain MRI using 3D convolutional autoencoders. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, pp. 5162–5165. http://dx.doi.org/10.1109/EMBC.2018.8513469.

Astaraki, M., Smedby, Ö., Wang, C., 2022. Prior-aware autoencoders for lung pathology segmentation. Med. Image Anal. 80, 102491. http://dx.doi.org/10.1016/j.media.2022.102491, URL https://www.sciencedirect.com/science/article/pii/S1361841522001384.

Atlason, H.E., Askell Love, M., Sigurdsson, S., Vilmundur Gudnason, M., Ellingsen, L.M., 2019. Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder. In: Angelini, E.D., Landman, B.A. (Eds.), In: Medical Imaging 2019: Image Processing, vol. 10949, SPIE, International Society for Optics and Photonics, p. 109491H. http://dx.doi.org/10.1117/12.2512953, URL https://doi.org/10.1117/12.2512953.

Baur, C., Graf, R., Wiestler, B., Albarqouni, S., Navab, N., 2020. SteGANomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain MRI. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2020, Springer International Publishing, Cham, pp. 718–727. http://dx.doi.org/10.1007/978-3-030-59713-9_69.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, Cham, pp. 161–169. http://dx.doi.org/10.1007/978-3-030-11723-8_16.

Bekkouch, I.E.I., Maksudov, B., Kiselev, S., Mustafaev, T., Vrtovec, T., Ibragimov, B., 2022. Multi-landmark environment analysis with reinforcement learning for pelvic abnormality detection and quantification. Med. Image Anal. 78, 102417. http://dx.doi.org/10.1016/j.media.2022.102417, URL https://www.sciencedirect.com/science/article/pii/S1361841522000688.

Brotto, D., Sorrentino, F., Cenedese, R., Avato, I., Bovo, R., Trevisi, P., Manara, R., 2021. Genetics of inner ear malformations: A review. Audiol. Res. 11 (4), 524–536. http://dx.doi.org/10.3390/audiolres11040047, URL https://www.mdpi.com/2039-4349/11/4/47.

Chakravorti, S., Noble, J.H., Gifford, R.H., Dawant, B.M., O'Connell, B., Wang, J., Labadie, R.F., 2019. Further evidence of the relationship between cochlear implant electrode positioning and hearing outcomes. Otol. Neurotol.: Off. Publ. Am. Otol. Soc., Am. Neurotol. Soc. Eur. Acad. Otol. Neurotol. 40 (5), 617. http://dx.doi.org/10.1097/MAO.0000000000002204.

Chen, X., You, S., Tezcan, K.C., Konukoglu, E., 2020. Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. 64, 101713. http://dx.doi.org/10.1016/j.media.2020.101713, URL https://www.sciencedirect.com/science/article/pii/S1361841520300773.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1995. Active shape models-their training and application. Comput. Vis. Image Underst. 61 (1), 38–59. http://dx.doi.org/10.1006/cviu.1995.1004.

Demarcy, T., Vandersteen, C., Guevara, N., Raffaelli, C., Gnansia, D., Ayache, N., Delingette, H., 2017. Automated analysis of human cochlea shape variability from segmented μCT images. Comput. Med. Imaging Graph. 59, 1–12. http://dx.doi.org/10.1016/j.compmedimag.2017.04.002, URL https://www.sciencedirect.com/science/article/pii/S0895611117300332.

Dhanasingh, A., Erpenbeck, D., Assadi, M.Z., Doyle, Ú., Roland, P., Hagr, A., Rompaey, V.V., de Heyning, P.V., 2021. A novel method of identifying inner ear malformation types by pattern recognition in the mid modiolar section. Sci. Rep. 11, http://dx.doi.org/10.1038/s41598-021-00330-6.

Dhanasingh, A.E., Weiss, N.M., Erhard, V., Altamimi, F., Roland, P., Hagr, A., Rompaey, V.V., de Heyning, P.V., 2022. A novel three-step process for the identification of inner ear malformation types. Laryngosc. Investig. Otolaryngol. http://dx.doi.org/10.1002/lio2.936, URL https://onlinelibrary.wiley.com/doi/10.1002/lio2.936.

Escudé, B., James, C., Deguine, O., Cochard, N., Eter, E., Fraysse, B., 2006. The size of the cochlea and predictions of insertion depth angles for cochlear implant electrodes. Audiol. Neurotol. 11 (Suppl. 1), 27–33.

Ghesu, F.C., Georgescu, B., Zheng, Y., Grbic, S., Maier, A., Hornegger, J., Comaniciu, D., 2019. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 176–189. http://dx.doi.org/10.1109/TPAMI.2017.2782687.

Gower, J.C., 1975. Generalized procrustes analysis. Psychometrika 40 (1), 33–51. http://dx.doi.org/10.1007/bf02291478.

Hainan, S., Jiang, M., Liu, Y., Lu, H., Liang, Z., 2019. The Detection of Non-Polypoid Colorectal Lesions Using the Texture Feature Extracted from Intact Colon Wall: A Pilot Study. SPIE-Intl Soc Optical Eng, p. 105. http://dx.doi.org/10.1117/12.2511823.

Hariri, S., Kind, M.C., Brunner, R.J., 2021. Extended isolation forest. IEEE Trans. Knowl. Data Eng. 33, 1479–1489. http://dx.doi.org/10.1109/TKDE.2019.2947676.

Korver, A.M., Smith, R.J., Van Camp, G., Schleiss, M.R., Bitner-Glindzicz, M.A., Lustig, L.R., Usami, S.-i., Boudewyns, A.N., 2017. Congenital hearing loss. Nat. Rev. Dis. Primers 3 (1), 1–17.

Leroy, G., Rueckert, D., Alansary, A., 2020. Communicative reinforcement learning agents for landmark detection in brain images. In: Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-Oncology. Springer, pp. 177–186. http://dx.doi.org/10.1007/978-3-030-66843-3_18.

Liu, F.T., Ting, K.M., Zhou, Z.H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data 6, http://dx.doi.org/10.1145/2133360.2133363.

López Diez, P., Juhl, K.A., Sundgaard, J.V., Diab, H., Margeta, J., Patou, F., Paulsen, R.R., 2022a. Deep reinforcement learning for detection of abnormal anatomies. In: Proceedings of the Northern Lights Deep Learning Workshop, vol. 3, UiT The Arctic University of Norway, http://dx.doi.org/10.7557/18.6280.

López Diez, P., Sørensen, K., Sundgaard, J.V., Diab, K., Margeta, J., Patou, F., Paulsen, R.R., 2022b. Deep reinforcement learning for detection of inner ear abnormal anatomy in computed tomography. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2022, Springer Nature Switzerland, Cham, pp. 697–706. http://dx.doi.org/10.1007/978-3-031-16437-8_67.

López Diez, P., Sundgaard, J.V., Patou, F., Margeta, J., Paulsen, R.R., 2021. Facial and cochlear nerves characterization using deep reinforcement learning for landmark detection. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2021, Springer International Publishing, Cham, pp. 519–528. http://dx.doi.org/10.1007/978-3-030-87202-1_50.

Lopez Pinaya, W.H., Vieira, S., Garcia-Dias, R., Mechelli, A., 2020. Chapter 11 - autoencoders. In: Mechelli, A., Vieira, S. (Eds.), Machine Learning. Academic Press, pp. 193–208. http://dx.doi.org/10.1016/B978-0-12-815739-8.00011-0, URL https://www.sciencedirect.com/science/article/pii/B9780128157398000110.

Margeta, J., Hussain, R., López Diez, P., Morgenstern, A., Demarcy, T., Wang, Z., Gnansia, D., Martinez Manzanera, O., Vandersteen, C., Delingette, H., Buechner, A., Lenarz, T., Patou, F., Guevara, N., 2022. A web-based automated image processing research platform for cochlear implantation-related studies. J. Clin. Med. 11 (22), http://dx.doi.org/10.3390/jcm11226640, URL https://www.mdpi.com/2077-0383/11/22/6640.

Martins, S.B., Telea, A.C., Falcão, A.X., 2020. Investigating the impact of supervoxel segmentation for unsupervised abnormal brain asymmetry detection. Comput. Med. Imaging Graph. 85, 101770. http://dx.doi.org/10.1016/j.compmedimag.2020.101770, URL https://www.sciencedirect.com/science/article/pii/S0895611120300720.

Masci, J., Meier, U., Cireşan, D., Schmidhuber, J., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (Eds.), Artificial Neural Networks and Machine Learning. ICANN 2011, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 52–59. http://dx.doi.org/10.1007/978-3-642-21735-7_7.

MONAI-Consortium, 2022. MONAI: Medical open network for AI. http://dx.doi.org/10.5281/zenodo.7459814.

Nazir, M., Shakil, S., Khurshid, K., 2021. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. Comput. Med. Imaging Graph. 91, 101940. http://dx.doi.org/10.1016/j.compmedimag.2021.101940, URL https://www.sciencedirect.com/science/article/pii/S0895611121000896.

Paludetti, G., Conti, G., Di Nardo, W., De Corso, E., Rolesi, R., Picciotti, P., Fetoni, A., 2012. Infant hearing loss: From diagnosis to therapy official report of XXI conference of Italian society of pediatric otorhinolaryngology. Acta Otorhinolaryngol. Italica 32 (6), 347.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Pawlowski, N., Lee, M.J., Rajchl, M., McDonagh, S.G., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., Zeiler, F.A., Digby, R., Coles, J.P., Rueckert, D., Menon, D.K., Newcombe, V.F.J., Glocker, B., 2018. Unsupervised lesion detection in brain CT using Bayesian convolutional autoencoders. In: 1st Conference on Medical Imaging with Deep Learning. MIDL 2018.

Pinaya, W.H.L., Graham, M.S., Gray, R., da Costa, P.F., Tudosiu, P.-D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., Werring, D., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2022a. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2022, Springer Nature Switzerland, Cham, pp. 705–714. http://dx.doi.org/10.1007/978-3-031-16452-1_67.

Pinaya, W.H., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2022b. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Med. Image Anal. 79, http://dx.doi.org/10.1016/j.media.2022.102475.

Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Med. Image Anal. 8 (3), 275–283. http://dx.doi.org/10.1016/j.media.2004.06.007, URL https://www.sciencedirect.com/science/article/pii/S1361841504000295. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003.

Radutoiu, A.T., Patou, F., Margeta, J., Paulsen, R.R., López Diez, P., 2022. Accurate localization of Inner Ear Regions of interests using deep reinforcement learning. In: Lian, C., Cao, X., Rekik, I., Xu, X., Cui, Z. (Eds.), Machine Learning in Medical Imaging. Springer Nature Switzerland, Cham, pp. 416–424. http://dx.doi.org/10.1007/978-3-031-21014-3_43.

Sajid, S., Hussain, S., Sarwar, A., 2019. Brain tumor detection and segmentation in MR images using deep learning. Arab. J. Sci. Eng. 44, 9249–9261. http://dx.doi.org/10.1007/s13369-019-03967-8.

Sato, D., Hanaoka, S., Nomura, Y., Takenaga, T., Miki, S., Yoshikawa, T., Hayashi, N., Abe, O., 2018. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes. In: Medical Imaging 2018: Computer-Aided Diagnosis, vol. 10575, SPIE, pp. 388–393. http://dx.doi.org/10.1117/12.2292276.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44. http://dx.doi.org/10.1016/j.media.2019.01.010, URL https://www.sciencedirect.com/science/article/pii/S1361841518302640.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (Eds.), Information Processing in Medical Imaging. Springer International Publishing, Cham, pp. 146–157. http://dx.doi.org/10.1007/978-3-319-59050-9_12.

Sennaroğlu, L., Bajin, M.D., 2017. Classification and current management of inner ear malformations. Balkan Med. J. 34, http://dx.doi.org/10.4274/balkanmedj.2017.0367.

Silva-Rodríguez, J., Naranjo, V., Dolz, J., 2022. Constrained unsupervised anomaly segmentation. Med. Image Anal. 80, http://dx.doi.org/10.1016/j.media.2022.102526.

Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J., 2020. An adversarial learning approach to medical image synthesis for lesion detection. IEEE J. Biomed. Health Inf. 24 (8), 2303–2314. http://dx.doi.org/10.1109/JBHI.2020.2964016.

Taboada-Crispi, A., Sahli, H., Orozco Monteagudo, M., Hernandez Pacheco, D., Falcon, A., 2009. Anomaly detection in medical image analysis. In: Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications. pp. 426–446. http://dx.doi.org/10.4018/978-1-60566-314-2.ch027, chapter 2.

Trier, P., Noe, K.Ø., Sørensen, M.S., Mosegaard, J., 2008. The visible ear surgery simulator. Stud. Health Technol. Inf. 132, 523.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imaging 20 (8), 677–688. http://dx.doi.org/10.1109/42.938237.

Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A., 2020. Attention guided anomaly localization in images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (Eds.), Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp. 485–503. http://dx.doi.org/10.1007/978-3-030-58520-4_29.

Vlontzos, A., Alansary, A., Kamnitsas, K., Rueckert, D., Kainz, B., 2019. Multiple landmark detection using multi-agent reinforcement learning. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2019, Springer International Publishing, Cham, pp. 262–270. http://dx.doi.org/10.1007/978-3-030-32251-9_29.

Wang, S., Zeng, Y., Yu, G., Cheng, Z., Liu, X., Zhou, S., Zhu, E., Kloft, M., Yin, J., Liao, Q., 2023. E3Outlier: A self-supervised framework for unsupervised deep outlier detection. IEEE Trans. Pattern Anal. Mach. Intell. 45, 2952–2969. http://dx.doi.org/10.1109/TPAMI.2022.3188763.

Wang, S., Zhu, Y., Lee, S., Elton, D.C., Shen, T.C., Tang, Y., Peng, Y., Lu, Z., Summers, R.M., 2022. Global-local attention network with multi-task uncertainty loss for abnormal lymph node detection in MR images. Med. Image Anal. 77, 102345. http://dx.doi.org/10.1016/j.media.2021.102345, URL https://www.sciencedirect.com/science/article/pii/S136184152100390X.

Watkins, C.J.C.H., 1989. Learning from Delayed Rewards (Ph.D. thesis). King's College, Cambridge, UK.

Welch, M.L., McIntosh, C., McNiven, A., Huang, S.H., Zhang, B.B., Wee, L., Traverso, A., O'Sullivan, B., Hoebers, F., Dekker, A., Jaffray, D.A., 2020. User-controlled pipelines for feature integration and head and neck radiation therapy outcome predictions. Phys. Med. 70, 145–152. http://dx.doi.org/10.1016/j.ejmp.2020.01.027.

Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C., 2022. Diffusion models for medical anomaly detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2022, Springer Nature Switzerland, Cham, pp. 35–45. http://dx.doi.org/10.1007/978-3-031-16452-1_4.

Xu, H., Pang, G., Wang, Y., Wang, Y., 2023a. Deep isolation forest for anomaly detection. IEEE Trans. Knowl. Data Eng. http://dx.doi.org/10.1109/TKDE.2023.3270293.

Xu, H., Wang, Y., Wei, J., Jian, S., Li, Y., Liu, N., 2023b. Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. URL http://arxiv.org/abs/2305.16114.

Yushkevich, P.A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31 (3), 1116–1128. http://dx.doi.org/10.1016/j.neuroimage.2006.01.015.

Zhou, S.K., Le, H.N., Luu, K., V Nguyen, H., Ayache, N., 2021. Deep reinforcement learning in medical imaging: A literature review. Med. Image Anal. 73, 102193. http://dx.doi.org/10.1016/j.media.2021.102193, URL https://www.sciencedirect.com/science/article/pii/S1361841521002395.

Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K., 2019. Unsupervised anomaly localization using variational auto-encoders. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2019, Springer International Publishing, Cham, pp. 289–297. http://dx.doi.org/10.1007/978-3-030-32251-9_32.

PAPER F

Unsupervised Classification of Congenital Inner Ear Malformations Using DeepDiffusion for Latent Space Representation

Authors Paula López Diez, Jan Margeta, Khassan Diab, François Patou, and Rasmus R. Paulsen.

Journal Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol 14224. Springer, Cham.

Year 2023

Status Published

DOI https://doi.org/10.1007/978-3-031-43904-9_63

Unsupervised Classification of Congenital Inner Ear Malformations Using DeepDiffusion for Latent Space Representation

Paula López Diez[1(✉)], Jan Margeta[3,4], Khassan Diab[5], François Patou[2], and Rasmus R. Paulsen[1]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
plodi@dtu.dk
[2] Oticon Medical, Research and Technology, Smørum, Denmark
[3] Oticon Medical, Research and Technology, Vallauris, France
[4] KardioMe, Research and Development, Nova Dubnica, Slovakia
[5] Tashkent International Clinic, Tashkent, Uzbekistan

Abstract. The identification of congenital inner ear malformations is a challenging task even for experienced clinicians. In this study, we present the first automated method for classifying congenital inner ear malformations. We generate 3D meshes of the cochlear structure in 364 normative and 107 abnormal anatomies using a segmentation model trained exclusively with normative anatomies. Given the sparsity and natural unbalance of such datasets, we use an unsupervised method for learning a feature representation of the 3D meshes using DeepDiffusion. In this approach, we use the PointNet architecture for the network-based unsupervised feature learning and combine it with the diffusion distance on a feature manifold. This unsupervised approach captures the variability of the different cochlear shapes and generates clusters in the latent space which faithfully represent the variability observed in the data. We report a mean average precision of 0.77 over the seven main pathological subgroups diagnosed by an ENT (Ear, Nose, and Throat) surgeon specialized in congenital inner ear malformations.

Keywords: Unsupervised · Classification · DeepDiffusion · Inner Ear

1 Introduction

Inner ear malformations are found in 20–30% of children with congenital hearing loss [1]. While the prevalence of bilateral congenital hearing loss is estimated to be 1.33 per 1000 live births in North America and Europe, it is much higher in sub-Saharan Africa (19 per 1,000 newborns) and South Asia (up to 24 per 1,000) [8]. Early detection of sensorineural hearing loss is crucial for appropriate intervention, such as cochlear implant therapy, which is prescribed to approximately

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 652–662, 2023.
https://doi.org/10.1007/978-3-031-43904-9_63

80,000 infants and toddlers annually worldwide [16]. Radiological examination is essential for an early diagnosis of congenital inner ear malformation, particularly when cochlear implant therapy is planned. However, detecting and classifying such malformations from standard imaging modalities is a complex task even for expert clinicians, and presents challenges during CI surgery [2]. Previous studies have proposed methods to classify congenital inner ear malformations based on explicit measurements and visual analysis of CT scans [5]. These methods are time-consuming and subject to clinician subjectivity. A suggested approach for the automated detection of inner ear malformation has relied on deep reinforcement learning trained for landmark location in normal anatomies based on an anomaly detection technique [9]. However, this method is only limited to the detection of a malformation but does not attempt to classify them.

Currently, supervised deep metric learning garners significant interest due to its exceptional efficacy in data clustering and pathology classification. Most of these approaches are fully supervised and use supervisory signals that model the training by creating tuples of labeled training data. These tuples are then used to optimize the intra-class distance of the different samples in the latent space, as has been done mostly for 2D images [15,20,21] and 2D representation of 3D images [4]. Several recent studies have demonstrated promising outcomes from unsupervised contrastive learning from natural images. However, their utility in the medical image domain is limited due to the high degree of inter-class similarity. Particularly in heterogeneous real clinical datasets in which the image quality and appearance can significantly impact the performance of such methods, rendering them less effective. In [22] an unsupervised strategy to learn medical visual representations by exploiting naturally occurring paired descriptive text in 2D images is proposed. Typically, in 3D images, an unsupervised low-dimensional representation is utilized for further clustering, as demonstrated in [14]. Nonetheless, such approaches are commonly developed using quite homogeneous datasets that are not representative of real-world applications and the diverse clinical settings in which they must operate.

Our objective is to develop a fully automated pipeline for the classification of inner ear malformations, utilizing a relatively large and unique dataset of such anomalies. The pipeline's design necessitates a profound comprehension of this data type and the congenital malformations themselves. Given the CT scans in this region are complex, and the images originate from diverse sources, we employ an unsupervised approach, uniquely based on the 3D shape of the cochlear structure. We have observed that the cochlear structure can be roughly but consistently segmented by a 3D-UNET model trained exclusively on normal cochlear anatomies. We then use these segmentations and adopt an entirely unsupervised approach, meaning the deep learning model is trained from scratch on these segmentations, and the class labels are not used for training. To map these shapes to an optimal latent space representation, we utilize DeepDiffusion, which combines the diffusion distance on a feature manifold with the feature learning of the encoder.

In this paper, we present the first automatic approach for the classification of congenital inner ear malformations. We use an unsupervised method to find the latent space representation of cochlear shapes, which allows for their further classification. We demonstrate that shapes from a segmentation model trained on normative cases, albeit imperfect, can be used to represent abnormalities. Moreover, our results indicate the potential for successfully applying this approach to other anatomies.

2 Data

Our dataset comprises a total of 485 clinical CT scans, consisting of 364 normal scans and 121 scans with various types of inner ear malformations. The distribution of inner ear scans for each type of malformation is shown in Fig. 1. We utilized the region-of-interest (ROI) extraction technique developed by [18], which involves selecting anatomical points of interest that are not part of the inner ear region to achieve a standardized and robust image orientation. To ensure consistency, all images were resampled to a spacing of 0.125 mm, and their intensities were normalized by scaling the 5^{th} and 95^{th} percentiles of the intensity distribution of each image to 0 and 1, respectively. Figure 1 also shows the data split used for training our model. We chose to use an approximate 50% split for abnormal cases, while the vast majority of normal cases, approximately 86%, were used for training. Other configurations were explored, including using only normal cases for training. However, it was demonstrated that while this approach may work for anomaly detection, it does not adequately categorize the different types of malformations.

3 Methods

3.1 Anatomical Representation

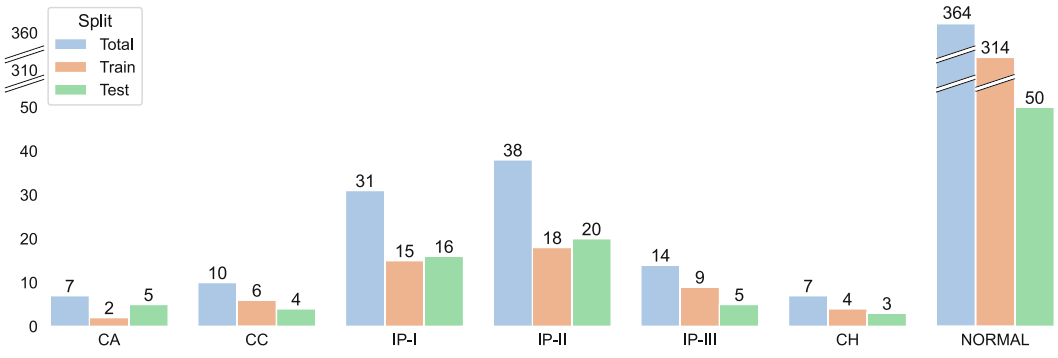


Fig. 1. Distribution of cases among the different classes and the split used for our approach. Cochlear aplasia (CA), common cavity (CC) incomplete partitioning type I, II, and III (IP-I, IP-II, IP-III), cochlear hypoplasia (CH), and normal.

Our aim is to find a parametrized shape that is representative of the anatomy of the patient. We decided to focus on the cochlear structure as it is the main

structure of interest when trying to identify a malformation in the inner ear. To obtain a 3D segmentation of this structure we use the 3D-UNET [19] presented in [10] which has been trained exclusively in normal anatomies (130 images from diverse imaging equipment) and built using MONAI [12]. Even though no abnormal anatomies have been used for training, given the high contrast between the soft tissue of the cochlear structure and the bony structure that surrounds it, the model still performs quite well to segment the abnormal cases. This can be seen in Fig. 2 where an example of each of the types of malformations used in this study and an anatomically normal case are shown. The largest connected component of the segmentation has been selected to generate the final 3D meshes.

An overview of our pipeline is presented in Fig. 3. Each 3D mesh obtained from a CT image is transformed into a 1024 point cloud using the Ohbuchi method [13]. Each shape is then normalized by centering its origin in its center of gravity and enclosing the shape within a unit sphere, resulting in the point cloud representation of the shape S. Before the shape S is fed to the encoder, the shape is augmented into shape $\hat{S}$ with a probability of 0.8. This augmentation consists of a random rotation with U(−5°,5°), an anisotropic scaling sampled from U(0.8,1), and a shearing and translation in each axes sampled from U(-0.2,0.2) for both actions.

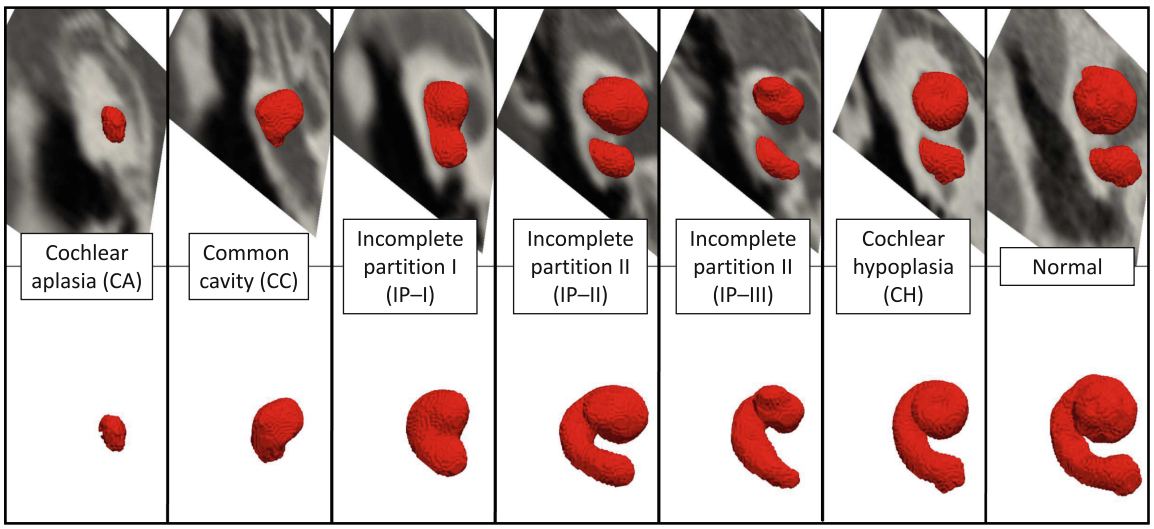


Fig. 2. Representative 3D-UNET segmentation meshes from each type of cochlear anatomy used in this study. Top row shows the 3D mesh with the original CT scan image; the bottom row shows exclusively the 3D mesh.

3.2 Deep Diffusion Algorithm

The DeepDiffusion (DD) algorithm [7] incorporates the manifold ranking [23] technique, which uses similarity diffusion on the manifold graph to learn a distance metric among the samples. The DD algorithm optimizes both the feature extraction and the embeddings produced by the encoder, which results in

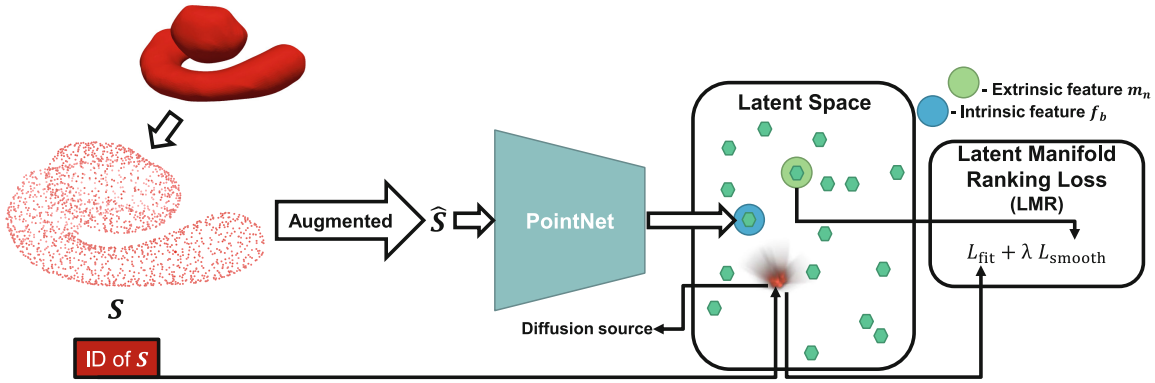


Fig. 3. Sketch of the DeepDiffusion used for latent space representation of the cochlear 3D meshes. The pointcloud extracted from the mesh is fed to the PointNet encoder which generates the corresponding latent feature which is optimized by minimizing the LMR loss so the encoder and the latent feature manifold are optimized for the comparison of data samples.

salient features in a continuous and smooth latent space. In this latent space, the Euclidean distance among the latent features approximates the diffusion distance on the latent feature manifold. The crux behind this algorithm is the latent manifold ranking loss (LMR) which is computed using both intrinsic and extrinsic features. The LMR consists of a fitting term, L_{fit}, a smoothing term, L_{smooth}, and a balancing term, λ.

$$\text{LMR} = \arg\min_{M,\theta} \; L_{\text{fit}} \pm \lambda L_{\text{smooth}} \tag{1}$$

Where θ characterizes the encoder and $M \in \mathbb{R}^{\text{N X P}}$ represents the latent feature manifold formed by the training samples, where N is the number of data samples and P is the output dimensions of the encoder. The extrinsic feature f is defined as the output of the encoder and has dimension P. M is initialized by stacking together the embeddings of the first forward pass through the encoder which has been randomly initialized as this has been shown to perform better than randomly initializing the weights of M itself as shown in [7].

Every training sample has its unique identification number (ID_b) which is used to specify a diffusion source y_b that is consistent throughout the training procedure. L_{fit} constrains the ranking vector r_b to being close to the diffusion source y_b, which is defined as the vector containing one-hot encoding of ID_b. The ranking vector is defined as $r_b = \text{softmax}(f_b M^T)$ and represents the probabilistic similarities between the feature f_b and all the intrinsic features contained in M. The fitting term is therefore defined as

$$L_{\text{fit}} = \sum_b \text{CrossEntropy}(r_b, y_b)$$

its minimization results in all the extrinsic features being embedded farther away from each other as they are being pulled toward their respective and unique

diffusion source vectors. The smoothing term is defined as

$$L_{\text{smooth}} = \sum_b \sum_n w_{bn} \text{Dissimilarity}(r_b, t_n) \tag{2}$$

where the dissimilarity operator is the Jensen-Shannon divergence [6] and $t_n = \text{softmax}(m_n M^T)$ being m_n the n^{th} row of the matrix M so that t_n contains the ranking score of the intrinsic feature m_n to all the intrinsic features. w_{bn} indicates the similarity between the extrinsic feature f_b and the neighboring intrinsic feature m_n and it is defined as:

$$w_{bn} = \begin{cases} f_b m_n^T, & \text{m}_n \in \text{kNN}(f_b) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Minimizing L_{smooth} pulls extrinsic features and their neighboring intrinsic features together which implies that an extrinsic feature is more likely to be projected onto the surface of the latent feature manifold of the intrinsic features when L_{smooth} is smaller.

3.3 Implementation

For our encoder, we use the PointNet [3] architecture which takes 1024 3D points as input, applies input and feature transformations, and then aggregates point features by max pooling to a feature of dimensionality 1024 which is then compressed into dimensionality 254 with two sets of fully connected layers. The network has been trained by using mini-batch (of size 8) gradient descent using the Adam optimizer with a learning rate of 10^{-8} and ReLU as the activation function. The DD algorithm is implemented in PyTorch [17] and the code used for this study is available at https://github.com/paulalopez10/Deep-Diffusion-Unsupervised-Classification-3D-Mesh. The models are trained on an NVIDIA GeForce RTX 3070 Laptop GPU with 8GB VRAM. The different hyper-parameters related to the approach have been explored and it has been empirically found for this specific configuration $\lambda = 0.6$ and $k = 10$ produce the best results that will be analyzed in the following section.

4 Results

We evaluate the classification performance of our pipeline by analyzing the embeddings generated by the trained encoder. To visualize the projection of the features of the test in 2D we use the U-MAP [11], as illustrated in Fig. 4. The U-MAP visualization demonstrates the clustering of different classes in the latent space. Furthermore, it is very interesting to notice how the latent space representation displays the anatomical changes of the anatomy where the more extreme types of malformations (CA and CC) are the most distant to the normative cochlear structures. The transition between the different classes shown in the latent space properly represents the pathological variations in this anatomy.

We have also included, in Fig. 4, the projection of the features projected in the 2D-PCA space defined by the training set, where both, training and testing, sets are included to show not only the clustering in this space but also the similar distribution of the different classes in both sets within the PCA projection.

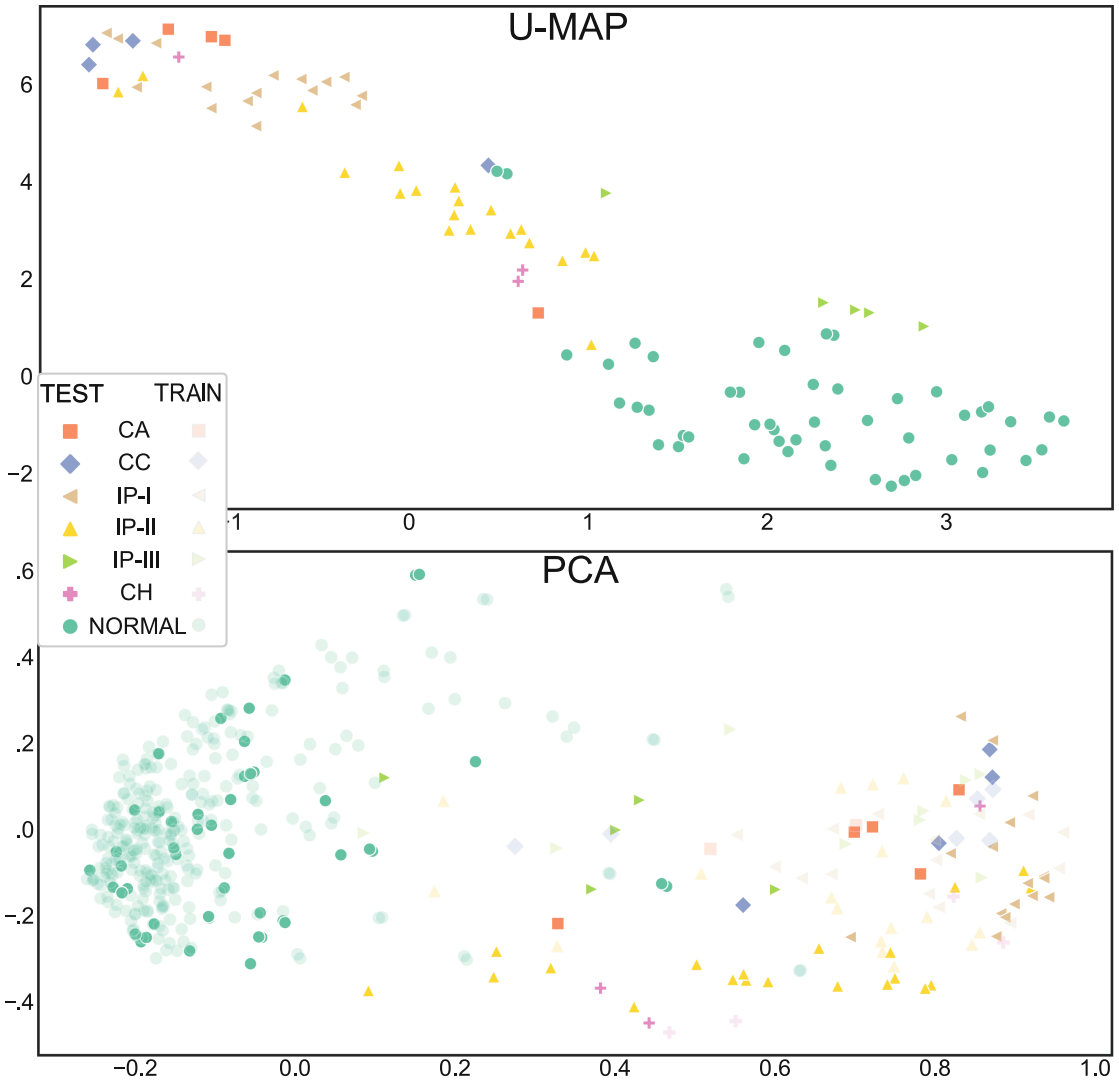


Fig. 4. Top: U-MAP representation of the test features where it can be observed how the different classes group together and how the anatomical variation is represented as there is a progression from the most abnormal cases towards fully normal cases. **Bottom**: Test and train features projected into the 2D PCA space defined by the training samples, where the classes are separated and show consistency between training and testing samples.

For a further analysis of the performance, we compute some evaluation metrics based on the pairwise cosine distance between samples that can be seen in Fig. 5 c). The average ROC and precision-recall curves for each of the classes can be seen in Fig. 5 a) and b). To calculate those, each test feature vector f_b is considered to be the centroid of a kNN(f_b) which consists in the k nearest

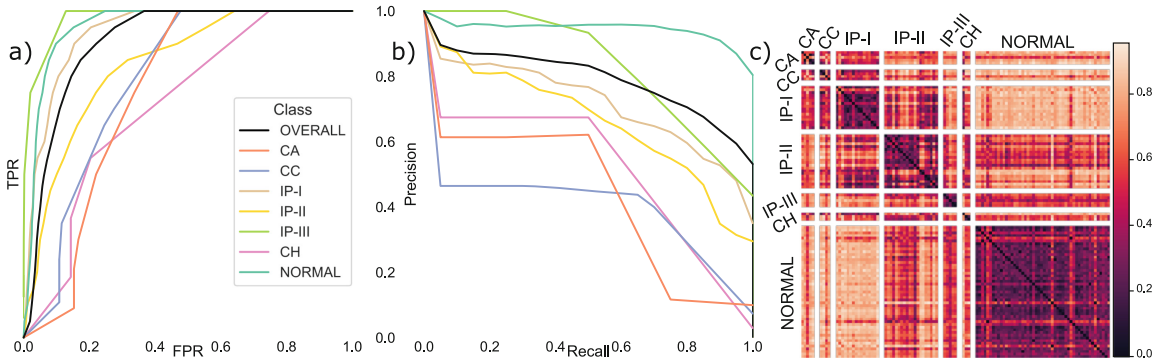


Fig. 5. Evaluation plots. **a)**Mean ROC curves for each class **b)**Mean Recall-Precision curve for each class **c)**Pairwise cosine distance between test embeddings used to evaluate the performance.

Table 1. Evaluation metrics reported in our experiment. ROC-Receiver operating characteristic, AUC- Area under the curve, AP-Average precision, PR - Precision-recall

	CA	CC	IP-I	IP-II	IP-III	CH	NORMAL	Overall
Max Accuracy	0.98	0.96	0.92	0.96	0.98	0.99	0.91	0.93
Mean Accuracy	0.73 ±0.25	0.77 ± 0.26	0.87 ± 0.03	0.77 ± 0.08	0.96 ± 0.03	0.69 ± 0.01	0.75 ± 0.05	0.78 ± 0.12
Max ROC-AUC	0.99	0.99	0.98	0.96	0.99	0.98	0.99	0.99
Mean ROC-AUC	0.75 ± 0.25	0.79 ± 0.25	0.94 ± 0.04	0.84 ± 0.10	0.98 ± 0.02	0.71 ± 0.08	0.95 ± 0.09	0.91 ± 0.13
Max AP	0.57	0.70	0.87	0.88	0.92	0.51	0.99	0.91
Mean AP	0.41 ± 0.42	0.38 ± 0.33	0.70 ± 0.23	0.65 ± 0.32	0.82 ± 0.25	0.50 ± 0.42	0.94 ± 0.12	0.77 ± 0.29
Max f1-score	0.67	0.57	0.67	0.88	0.86	0.67	0.91	0.75
Max PR-AUC	0.63	0.80	0.84	0.85	0.91	0.71	0.95	0.88
Mean PR-AUC	0.40 ± 0.26	0.37 ± 0.25	0.67 ± 0.04	0.62 ± 0.11	0.78 ± 0.02	0.48 ± 0.08	0.90 ± 0.10	0.74 ± 0.24

features from other samples using the cosine pairwise distance shown in Fig. 5 c). We vary k until all the features from the corresponding class are within the cluster and compute the precision and false positive rate per the different recall steps, the shown results are the average among each class and overall. With the same procedure, different evaluation metrics have been obtained and are shown in Table 1. These metrics encompass the area under the curve (AUC) for the curves shown in Fig. 5 a) and b), both for the average curve and the optimal curve for each class. Furthermore, the maximum and average accuracy has been computed together with the maximum f1-score. Considering the dataset's significant class imbalance, these metrics provide a comprehensive assessment of the performance achieved. Finally, the mean average precision is also included in the table together with the optimal one for each class. The optimal or maximum value of each metric corresponds to when the optimal sample within our test features distribution is being evaluated as the centroid of its own class and the

mean values are the average over all the samples. We can observe how a bigger variance is obtained for the classes that contain a few examples as it is expected, given the nature and distribution of our dataset shown in Fig. 1.

5 Conclusion

We have presented the first approach for the automatic classification of congenital inner ear malformations. We show how using the 3D shape information of the cochlea obtained with a model only trained in normative anatomies is enough to classify the malformations and reduces the influence of the image's source, which is crucial in a clinical application setting.

Our method shows a mean average precision of 0.77 with a mean ROC-AUC of 0.91, indicating its effectiveness in classifying inner ear malformations. Furthermore, the representation of the different cases in the latent space shows spatial relation between classes, which is correlated with the anatomical appearance of the different malformations. These promising results pave the way towards assisting clinicians in the challenging assessment of congenital inner ear malformations potentially leading to improved patient outcome of cochlear surgery.

References

1. Brotto, D., et al.: Genetics of inner ear malformations: a review. Audiol. Res. **11**(4), 524–536 (2021). https://doi.org/10.3390/audiolres11040047
2. Chakravorti, S., et al.: Further evidence of the relationship between cochlear implant electrode positioning and hearing outcomes. Otol. Neurotol. **40**(5), 617–624 (2019). https://doi.org/10.1097/MAO.0000000000002204
3. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 77–85 (2017). https://doi.org/10.1109/CVPR.2017.16
4. Chen, X., Wang, W., Jiang, Y., Qian, X.: A dual-transformation with contrastive learning framework for lymph node metastasis prediction in pancreatic cancer. Med. Image Anal. **85**, 102753 (2023). https://doi.org/10.1016/j.media.2023.102753, https://www.sciencedirect.com/science/article/pii/S1361841523000142
5. Dhanasingh, A.E., et al.: A novel three-step process for the identification of inner ear malformation types. Laryngoscope Investigative Otolaryngology (2022). https://doi.org/10.1002/lio2.936, https://onlinelibrary.wiley.com/doi/10.1002/lio2.936
6. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings, pp. 31- (2004). https://doi.org/10.1109/ISIT.2004.1365067
7. Furuya, T., Ohbuchi, R.: Deepdiffusion: unsupervised learning of retrieval-adapted representations via diffusion-based ranking on latent feature manifold. IEEE Access **10**, 116287–116301 (2022). https://doi.org/10.1109/ACCESS.2022.3218909

8. Korver, A.M., et al.: Congenital hearing loss. Nature Rev. Disease Primers **3**(1), 1–17 (2017)
9. López Diez, P., et al.: Deep reinforcement learning for detection of inner ear abnormal anatomy in computed tomography. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022, pp. 697–706. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_67
10. Margeta, J., et al.: A web-based automated image processing research platform for cochlear implantation-related studies. J. Clin. Med. **11**(22) (2022). https://doi.org/10.3390/jcm11226640, https://www.mdpi.com/2077-0383/11/22/6640
11. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861
12. MONAI-Consortium: Monai: Medical open network for AI (2022). https://doi.org/10.5281/zenodo.7459814
13. Ohbuchi, R., Minamitani, T., Takei, T.: Shape-similarity search of 3D models by using enhanced shape functions. Int. J. Comput. Appl. Technol. **23**(2–4), 70–85 (2005)
14. Onga, Y., Fujiyama, S., Arai, H., Chayama, Y., Iyatomi, H., Oishi, K.: Efficient feature embedding of 3d brain mri images for content-based image retrieval with deep metric learning. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 3764–3769. IEEE (2019)
15. Pal, A., et al.: Deep metric learning for cervical image classification. IEEE Access **9**, 53266–53275 (2021)
16. Paludetti, G., et al.: Infant hearing loss: from diagnosis to therapy official report of xxi conference of Italian society of pediatric otorhinolaryngology. Acta Otorhinolaryngol. Italica **32**(6), 347 (2012)
17. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
18. Radutoiu, A.T., Patou, F., Margeta, J., Paulsen, R.R., López Diez, P.: Accurate localization of inner ear regions of interests using deep reinforcement learning. In: Lian, C., Cao, X., Rekik, I., Xu, X., Cui, Z. (eds.) Machine Learning in Medical Imaging. pp. 416–424. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-21014-3_43
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, pp. 234–241. Springer International Publishing, Cham (2015)
20. Sundgaard, J.V., et al.: Deep metric learning for otitis media classification. Med. Image Anal. **71**, 102034 (2021). https://doi.org/10.1016/j.media.2021.102034, https://www.sciencedirect.com/science/article/pii/S1361841521000803
21. Zhang, Y., Luo, L., Dou, Q., Heng, P.A.: Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. Med. Image Anal. **86**, 102772 (2023). https://doi.org/10.1016/j.media.2023.102772, https://www.sciencedirect.com/science/article/pii/S1361841523000336

22. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 2–25. PMLR (2022). https://proceedings.mlr.press/v182/zhang22a.html
23. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems. vol. 16. MIT Press (2003). https://proceedings.neurips.cc/paper/2003/file/2c3ddf4bf13852db711dd1901fb517fa-Paper.pdf

PAPER G

Peridevice leaks following left atrial appendage occlusion - analysis with morphology descriptive centerlines and explainable graph attention network

Authors Paula López Diez, Jan Margeta, Javier Gómez-Herrero, Davorka Lulic, Yannick Willemen, Klaus F. Kofoed, Ole De Backer, and Rasmus R. Paulsen.

Journal Proc. of 15th International Workshop, STACOM 2024, Held in Conjunction with MICCAI, 2024. Lecture Notes in Computer Science. Springer, Cham.

Year 2024

Status In proceedings

Peridevice leaks following left atrial appendage occlusion - analysis with morphology descriptive centerlines and explainable graph attention network

Paula López Diez[1], Jan Margeta[2], Javier Gómez-Herrero[5], Davorka Lulic[3], Yannick Willemen[3], Klaus F. Kofoed[3,4], Ole De Backer[3], and Rasmus R. Paulsen[1]

[1] DTU Compute, Technical University of Denmark, Denmark
[2] Research & Development, KardioMe, Slovakia
[3] Department of Cardiology, Copenhagen University Hospital - Rigshospitalet, Denmark
[4] Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[5] Servicio de Cardiología, Hospital Clínico Universitario de Valladolid, Spain

Abstract. Peridevice leaks (PDLs) following left atrial appendage occlusion can negate the protective effect of the procedure and are a challenge in clinical cardiology. In this study, we investigate how the morphology of the anatomy relevant to the procedure interacts with the risk of PDLs using pre-operative CT images. We construct 3D models of the anatomy involved in the transcatheter procedure from a dataset of 125 patients who underwent left atrial appendage occlusion, integrating an anatomical centerline that simulates the catheter trajectory during device placement and that is complemented with morphological features. Given the complex and variable nature of this anatomy, we utilize Graph Attention Networks to analyze these models. This architecture enables us to encode morphological features into a graph structure, capturing the intricate spatial relationships and dependencies. We predict the likelihood of potential PDLs and identify key morphological descriptors contributing to these predictions through attention scores. This method provides a standardized representation of the anatomy involved in the procedure, offering insights into the anatomical factors influencing PDL risk and aiding in pre-operative planning.

Keywords: LAAO · GAT · centerline · leak · explainability.

1 Introduction

The left atrial appendage (LAA) is a small pouch in the left atrium of the heart, often implicated in thrombus formation, particularly in patients with atrial fibrillation (AF), in which around 90% of intracardiac thrombi are found in the LAA [20]. To mitigate stroke risk, left atrial appendage occlusion (LAAO) is

performed to seal off the LAA and prevent thromboembolic events in AF patients. This transcatheter intervention involves placing a device that occludes the LAA from the LA. However, peridevice leaks (PDLs) can occur, where gaps between the occluder and the LAA wall allow blood to bypass the device.

PDLs are significant as they can negate the protective effect of LAAO. Recent studies have documented a relationship between PDLs and subsequent thromboembolic events[1]. These leaks arise from incomplete endothelialization, suboptimal device placement, or anatomical variations in the LAA. It has also been shown that CT is the most sensitive method to evaluate leaks because it provides a detailed perspective of the shape, size, and mechanism of the PDL[7]. Managing PDLs involves monitoring, pharmacological therapy, and sometimes secondary interventions. Predicting PDL occurrence and understanding their morphology remain challenging due to the complex geometries and dynamic nature of the LAA [3]. In this study, we want to explore to which degree the morphology of the anatomy is related to the potential risk of leakage. It is known that some anatomies are more difficult for the operator. We aim to find a standardized representation of the complex anatomy that captures the challenges involved in the procedure.

Given the high complexity and inter-patient variability of the LAA anatomy, standardization presents a significant challenge from an image analysis standpoint. However, many different methods have been developed in the past to optimize the LAAO planning. Some focus on the risk of thromboembolic events based on the morphology of the appendage itself [5], sometimes analyzing specific anatomical features of the LAA anatomy or the combination with hemodynamic properties[13, 15]. Some research also focuses on helping clinicians choose the optimal device for a specific anatomy using both the morphology and computational fluid dynamics on the LA anatomy [9, 2]. PDLs are a complex and multifactorial problem that has not yet been investigated in depth. It has been shown that possible predictors of PDLs are the length of the procedure or the number of attempts [8]. Even though these can be used to select patients for a closer follow-up, it is not applicable for pre-procedural planning. We focus on the analysis of how the anatomical centerline that mimics the expected catheter trajectory can help us detect morphologies with a higher risk of PDL after undergoing LAAO.

Recent advancements in machine learning, particularly graph attention networks (GATs), offer promising approaches for processing graph-structured data, capturing the intricate relationships and spatial dependencies [10]. Additionally, incorporating explainability features into these networks allows for a better understanding of the morphological descriptors contributing to a decision of the trained classifier.

In this paper, we present a novel method for characterizing the anatomy involved in the LAAO by extracting the morphological centerline that simulates the trajectory of the catheter for device placement. We build a graph based on this centerline and encode other relevant features of the morphology. Furthermore, we train a GAT model to predict the PDL likelihood given a pre-operative

anatomy and use the attention scores to explain which parts of the graph had a higher influence on the predicted outcome.

2 Data

The dataset used for this study consists of 125 patients who have undergone LAAO. For each patient, pre-operative and post-operative CT images have been collected. The pre-operative CT images were used to extract the graph characterizing the anatomy. The post-operative CT images were analyzed by two expert annotators to determine whether a leak was present in the LAA by examining whether the contrast agent reached the distal area of the LAA. Furthermore, the different types of leaks have been classified according to their nature. The population distribution and examples of each type of leak are shown in Figure 1.

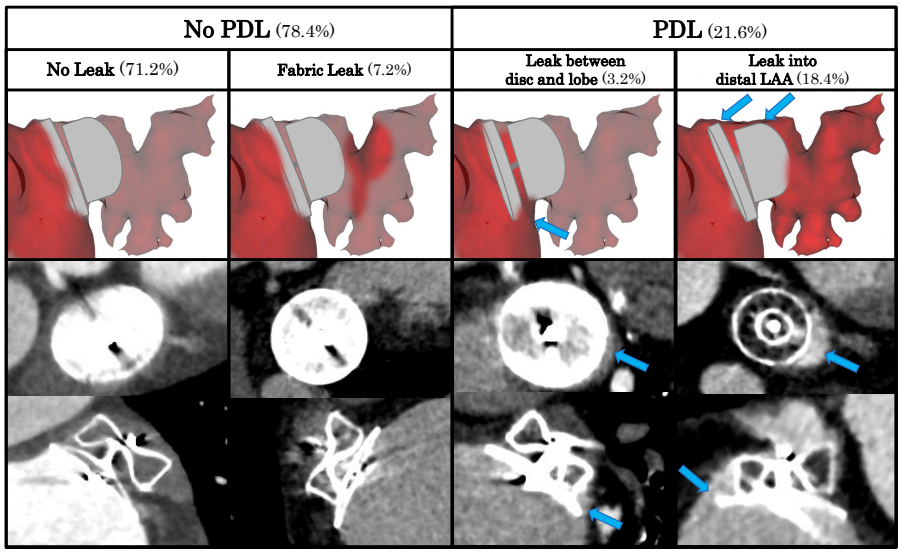


Fig. 1. Four classes defined based on the mechanisms of leaks after LAAO with a disc-and-lobe occluder. Prevalence of each class over the 125 subjects used for this study. The top row shows an illustration for conceptual understanding in 3D, underneath 2 views of an example CT image from our study. Blue arrows indicate the region in which the gap associated to the PDL can be observed.

All patients were treated by the same medical team, utilizing three different types of devices. Two of the devices are disc-and-lobe occluders: Amplatzer Amulet™(with an incidence of 32.8%) and Omega™(with an incidence of 36.8%); and the third type is a plug-in type: Watchman™(with an incidence of 30.4%).

Leaks were classified into PDL and non-PDLs. A PDL is indicated by a visible gap between the device and the anatomy, even if the gap is only in the disc part and the lobe is sealed. Non-PDL cases include those with no patency or with patency but no visible gap; the latter are typically fabric leaks or micro leaks expected to seal over time [14]. Usually, in these cases the CT has been obtained before the process of endothelialization is completed and therefore is not considered a PDL in our dataset. It should be noted that, given the nature of the occluders, only the disc-and-lobe devices can have a leak between the disc and the lobe whereas fabric leaks or micro leaks are possible for all the devices. An example of each of these categories both in an anatomical illustration and a CT image can be observed in Figure 1.

3 Methods

3.1 Graph generation

To construct a graph that characterizes the anatomy in a way that is representative of the procedure's complexity, we have decided to extract the centerline of the anatomical path used for the device placement. There is a high degree of heterogeneity among patients regarding the morphology of the access to the LAA, which can be observed in Figure 2, which may influence the ability to achieve a coaxial alignment of the occluder with the LAA neck[11]. The vascular access to implant the device is usually done through the femoral vein, continues through the inferior vena cava (IVC), and enters the heart through the right atrium (RA). Then, a transseptal puncture is performed in the fossa ovalis (FO) to access the LA and, finally, the appendage is reached [17]. Given the importance of reaching the appendage with the desired angle and coaxiality that allows for optimal placement and the limited degrees of freedom given by the catheter; the pre-operative planning involves an exhaustive analysis of the anatomy, often using CT images to analyze its morphology and the potential risks associated with a specific patient.

We use TotalSegmentator [19] to obtain the segmentation of the LA, IVC, and RA from the pre-operative CT image. However, due to the highly variable shape of the LAA, which TotalSegmentator locates but does not provide a precise segmentation for, we employ the method proposed in [16]. This method learns the neural unsigned distance field and generates much more precise and detailed segmentations of the LAA.

Once all the relevant anatomies are extracted from the CT, we obtain the joint mesh by iso-surface extraction and proceed to generate a standardized opening that resembles the transseptal puncture. To do so, we locate the closest point between the RA and the LA, which is the thinnest point of the FO, and generate an opening of 5mm radius. After this, we extract the centerline with VMTK [4]. We use the geodesically furthest point in the IVC with respect to the RA as the initial point and the geodesically furthest point in the LAA with respect to the LA as the endpoint. After extracting the centerline, we trim it so it starts 1 cm inside the IVC. Additionally, we fit a third-degree B-spline to

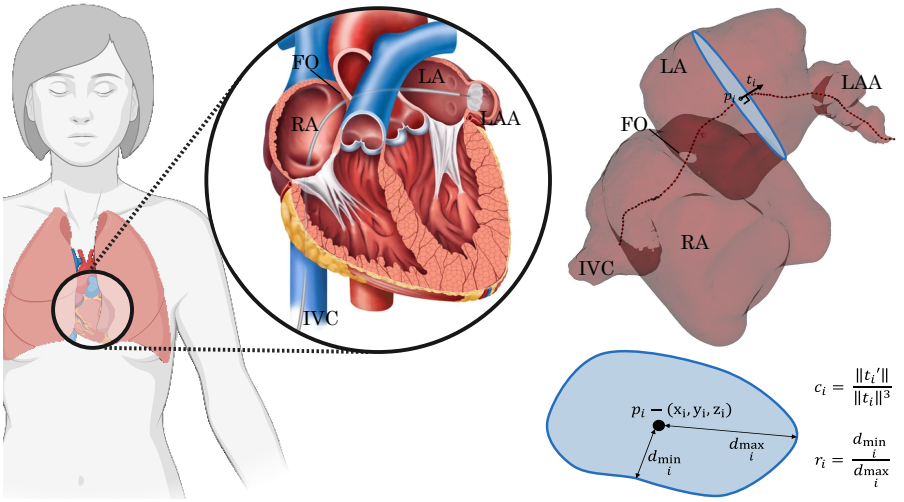


Fig. 2. Left: Illustration of the heart including an exemplified trajectory of the catheter to release the device in the neck of the LAA during a transcatheter LAAO procedure. **Right:** Anatomy and computed centerline extracted from the CT image including a sketch of the cross-sectional plane used to obtain $d_{\min}_i$ and $d_{\min}_i$, together with the equations used to calculate c_i and r_i.

this centerline and sample 100 points, which serve as the nodes of our graph. In Figure 2, we can observe an example of the final centerline obtained with this method. To further characterize the anatomy and provide descriptive features to the nodes, we compute the tangent t_i for each point p_i. Using the tangent, we can obtain the estimated curvature in each point c_i using the equation shown in Figure 2. Finally, we also extract the cross-sectional plane for each point p_i defined by the tangent t_i. From this plane, we obtain the shortest distance between the surface and the center point $d_{\min}_i$ and the largest one $d_{\max}_i$, as illustrated in Figure 2. The last descriptive parameter associated with a point p_i is the ratio between the shortest and the largest distance in this cross-sectional plane, which we define as r_i and for which the computational details are shown in Figure 2.

Figure 3 displays the 125 different graphs from our population, demonstrating the variability within our dataset. To explore this variability and see if there was any clear correlation between the extracted graphs and the presence of PDLs we project them into the 2D PCA space using the 100x7 features matrix of each graph, defined by the number of points and the corresponding 7 features associated with each of them. These features include the spatial coordinates x_i, y_i and z_i as well as derived features $d_{\min}_i$, $d_{\max}_i$, c_i, and r_i. We can see in Figure 3 that both distributions are similar with no clusters being formed. After a careful statistical analysis and exploring different fundamental methods to find potential

differences between the two groups, results ruled out significant differences and we moved towards more advanced methods.

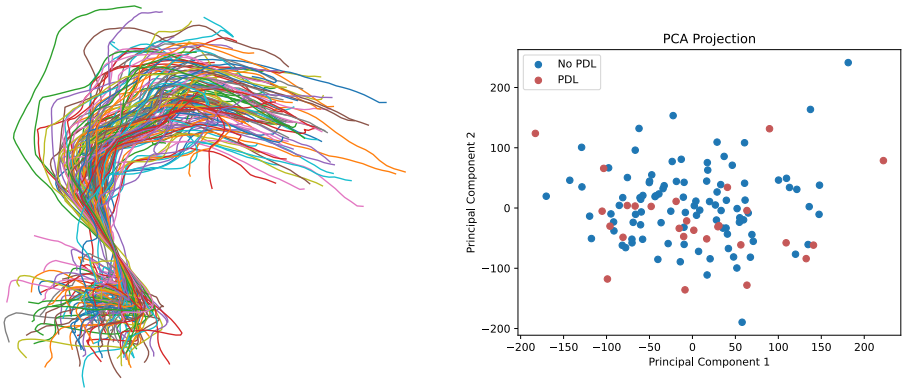


Fig. 3. Left: Display of the 125 computed centerlines. **Right:** PCA distribution of the 125 subjects (color-coded PDL/no PDL) using 7 features for each of the 100 points.

3.2 Graph Attention Network

For each subject, we can construct a structured graph using the method described in the previous section with 100 nodes and 7 features. All graphs are bidirectional and centered with respect to the FO.

GAT [18] is a type of graph neural network model that encodes edges into an attention mechanism and uses eigenvectors and eigenvalues of each node as positional embeddings for local structures. Due to the relatively simple structure of our graphs, we designed a streamlined architecture for graph-level predictions (Figure 4). This architecture consists of 2 GAT convolutions followed by global mean pooling and a fully connected layer. We utilize multi-head attention with 6 heads that are concatenated.

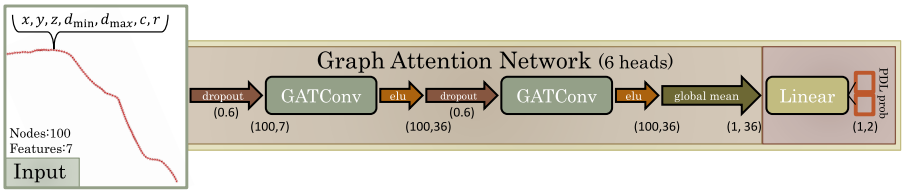


Fig. 4. Sketch of the GAT architecture used to classify the anatomical descriptive graphs of the centerline. From left to right, we present the input, the layers of the GAT architecture, and finally, the fully connected layers that map the output.

One advantage of using a multi-head GAT is its capability to generate explanations based on attention coefficients at the edge level. After processing a graph, we aggregate attention scores across layers and heads by selecting the maximum, generating a new graph where each edge contains its corresponding final attention score.

3.3 Implementation

We split our dataset into 70% for training and 30% for testing, maintaining the distribution of the different classes shown in Figure 1. To address the class imbalance, we use a weighted version of the negative log-likelihood loss as our loss function.

The code used for generating graphs from the segmentation images, as well as for training, testing, and explaining the GAT model, is available at `https://github.com/paulalopez10/GAT-graph_classification`, where you can find all the implementation details. We trained the model on an NVIDIA GeForce RTX 4090 with a training time of 6 hours. The code is built on PyTorch [12] and Pytorch Geometric[6].

4 Results and Discussion

We evaluate the performance of the GAT model in the graph-level binary classification task of predicting PDL or No PDL. The results, presented as a confusion matrix, are displayed in Table 1.

Table 1: Confusion matrix of the evaluation in the test set including the percentage of the corresponding true label.

		Predicted Label		
		No PDL	**PDL**	Total
True Label	**No PDL**	29(100%)	0(0%)	29
	PDL	5(62.5%)	3 (37.5%)	8
	Total	34	3	37

Given the multifactorial nature of the problem we are trying to tackle, we consider the results extremely positive. We observe that the model has been able to identify underlying patterns in the morphology described by our constructed graphs, which might indicate a higher chance of presenting a PDL after LAAO. In other words, the model has found some indication of the complexity of the procedure given a specific anatomy.

To analyze what the model is basing its predictions on, we choose the samples with the predicted highest probability of PDL and No PDL, respectively. Both cases are correctly classified according to our post-operative analysis. In Figure 5, we show the attention scores in the corresponding graphs together with an overlay of the given anatomy. We can observe that the model is focusing in

the region between the IVC, RA, and FO, where the highest change in attention scores is present. It is shown that this area can exhibit a higher level of curvature compared with other regions of the centerline. This abrupt change is sometimes present if the angle of the IVC and the FO with respect to the LAA deviates significantly from 180°. Given the nature of the procedure, it appears that the high curvature in that area could be linked with more difficult alignment, making the achievement of the desired coaxiality between the device and the LAA more challenging.

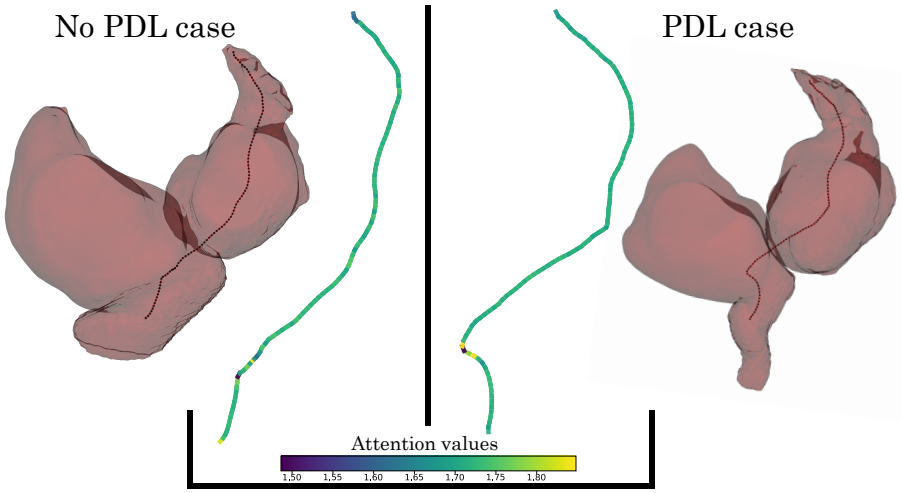


Fig. 5. Display of the attention scores associated with the edges of each graph generated during inference in two samples from the test set with our multi-headed GAT. On the left the case of the test set with the highest probability of No PDL (correctly classified), on the right the case with the highest probability of PDL (correctly classified).

5 Conclusion

We have presented a new method for characterizing the anatomy involved in the transcatheter LAAO procedure. We demonstrate that our standardized and automatic method, which characterizes the anatomy by generating a morphology-aware graph based on the centerline that mimics the catheter trajectory, successfully captures key elements to determine the complexity of the procedure.

We show how a graph attention network can be trained on these structures to predict PDLs, using the attention scores to highlight the regions of our graph that have a greater impact on the algorithm's decision. We observe that the curvature before the FO is the area of highest interest to our model for mapping the probability of a potential PDL. This allows us to observe underlying trends

that would otherwise be very difficult to analyze given the high complexity of the studied anatomy and the significant inter-personal morphological variability. This paves the way toward characterizing the difficulty on a patient-specific basis and predicting potential procedure outcomes.

References

1. Afzal, M.R., Gabriels, J.K., Jackson, G.G., Chen, L., Buck, B., Campbell, S., Sabin, D.F., Goldner, B., Ismail, H., Liu, C.F., Patel, A., Beldner, S., Daoud, E.G., Hummel, J.D., Ellis, C.R.: Temporal changes and clinical implications of delayed peridevice leak following left atrial appendage closure. JACC: Clinical Electrophysiology **8**(1), 15–25 (2022). https://doi.org/https://doi.org/10.1016/j.jacep.2021.06.018
2. Aguado, A.M., Olivares, A.L., Yaguë, C., Silva, E., Nuñez-Garciá, M., Álvaro Fernandez-Quilez, Mill, J., Genua, I., Arzamendi, D., Potter, T.D., Freixa, X., Camara, O.: In silico optimization of left atrial appendage occluder implantation using interactive and modeling tools. Frontiers in Physiology **10** (2019). https://doi.org/10.3389/fphys.2019.00237
3. Alkhouli, M., Backer, O.D., Ellis, C.R., Nielsen-Kudsk, J.E., Sievert, H., Natale, A., Lakkireddy, D., Holmes, D.R.: Peridevice leak after left atrial appendage occlusion: Incidence, mechanisms, clinical impact, and management (3 2023). https://doi.org/10.1016/j.jcin.2022.12.006
4. Antiga, L., Piccinelli, M., Botti, L., Ene-Iordache, B., Remuzzi, A., Steinman, D.A.: An image-based modeling framework for patient-specific computational hemodynamics (2008). https://doi.org/10.1007/s11517-008-0420-1
5. Ferez, X.M., Mill, J., Juhl, K.A., Acebes, C., Iriart, X., Legghe, B., Cochet, H., Backer, O.D., Paulsen, R.R., Camara, O.: Deep learning framework for real-time estimation of in-silico thrombotic risk indices in the left atrial appendage. Frontiers in Physiology **12** (6 2021). https://doi.org/10.3389/fphys.2021.694945
6. Fey, M., Lenssen, J.E.: Fast graph representation learning with pytorch geometric (2019)
7. Korsholm, K., Jensen, J.M., Nørgaard, B.L., Samaras, A., Saw, J., Berti, S., Tzikas, A., Nielsen-Kudsk, J.E.: Peridevice leak following amplatzer left atrial appendage occlusion: Cardiac computed tomography classification and clinical outcomes. JACC: Cardiovascular Interventions **14**(1), 83–93 (2021). https://doi.org/https://doi.org/10.1016/j.jcin.2020.10.034
8. Lakkireddy, D., Nielsen-Kudsk, J.E., Windecker, S., Thaler, D., Price, M.J., Gambhir, A., Gupta, N., Koulogiannis, K., Marcoff, L., Mediratta, A., Anderson, J.A., Gage, R., Ellis, C.R.: Mechanisms, predictors, and evolution of severe peri-device leaks with two different left atrial appendage occluders. EP Europace **25**(9), euad237 (08 2023). https://doi.org/10.1093/europace/euad237, `https://doi.org/10.1093/europace/euad237`
9. Mill, J., Montoliu, H., Moustafa, A.H., Olivares, A.L., Albors, C., Aguado, A., Medina, E., Ceresa, M., Freixa, X., Arzamendi, D., Cochet, H., Camara, O.: Domain expert evaluation of advanced visual computing solutions for the planning of left atrial appendage occluder interventions . https://doi.org/10.1101/2022.04.11.22273553, `https://doi.org/10.1101/2022.04.11.22273553`
10. Nerrise, F., Zhao, Q., Poston, K.L., Pohl, K.M., Adeli, E.: An explainable geometric-weighted graph attention network for identifying functional networks associated with gait impairment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 723–733. Springer (2023)

11. Pastormerlo, L.E., De Caterina, A.R., Esposito, A., Korsholm, K., Berti, S.: State-of-the-art of transcatheter left atrial appendage occlusion. Journal of Clinical Medicine **13**(4) (2024). https://doi.org/10.3390/jcm13040939
12. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
13. Saiz-Vivó, M., Mill, J., Harrison, J., Jimenez-Pérez, G., Legghe, B., Iriart, X., Cochet, H., Piella, G., Sermesant, M., Camara, O.: Unsupervised machine learning exploration of morphological and haemodynamic indices to predict thrombus formation in the left atrial appendage. vol. 13593 LNCS, pp. 200–210. Springer Science and Business Media Deutschland GmbH (2022). https://doi.org/10.1007/978-3-031-23443-9_19
14. Saw, J., Fahmy, P., DeJong, P., Lempereur, M., Spencer, R., Tsang, M., Gin, K., Jue, J., Mayo, J., McLaughlin, P., Nicolaou, S.: Cardiac CT angiography for device surveillance after endovascular left atrial appendage closure. European Heart Journal - Cardiovascular Imaging **16**(11), 1198–1206 (04 2015). https://doi.org/10.1093/ehjci/jev067, https://doi.org/10.1093/ehjci/jev067
15. Smine, Z., Melidoro, P., Qureshi, A., Longobardi, S., Williams, S.E., Aslanidi, O., Vecchi, A.D.: Global sensitivity analysis of thrombus formation in the left atrial appendage of atrial fibrillation patients. vol. 14507 LNCS, pp. 55–65. Springer Science and Business Media Deutschland GmbH (2024). https://doi.org/10.1007/978-3-031-52448-6_6
16. Sørensen, K., Camara, O., Backer, O.d., Kofoed, K.F., Paulsen, R.R.: Nudf: Neural unsigned distance fields for high resolution 3d medical image segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2022). https://doi.org/10.1109/ISBI52829.2022.9761610
17. Tzikas, A., Gafoor, S., Meerkin, D., Freixa, X., Cruz-González, I., Lewalter, T., Saw, J., Berti, S., Nielsen-Kudsk, J.E., Ibrahim, R., Lakkireddy, D., Paul, V., Arzamendi, D., Nietlispach, F., Worthley, S.G., Hildick-Smith, D., Thambo, J.B., Tondo, C., Aminian, A., Kalarus, Z., Schmidt, B., Søndergaard, L., Kefer, J., Meier, B., Park, J.W., Sievert, H., Omran, H.: Left atrial appendage occlusion with the amplatzer amulet device: an expert consensus step-by-step approach. EuroIntervention **11**(13), 1512–1521 (2016). https://doi.org/10.4244/EIJV11I13A292
18. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks (2018)
19. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
20. Wunderlich, N.C., Beigel, R., Swaans, M.J., Ho, S.Y., Siegel, R.J.: Percutaneous interventions for left atrial appendage exclusion: Options, assessment, and imaging using 2d and 3d echocardiography. JACC: Cardiovascular Imaging **8**(4), 472–488 (2015). https://doi.org/10.1016/j.jcmg.2015.02.002

$\sqrt{17}$
$\Omega \int \delta e^{i\pi} = -1$
{2.7182818284}
ϑφερτυθιοπσδφγηξκλ
χ^2
$\sum$