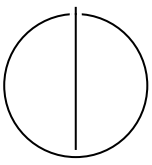
SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

AI-Based Cochlear Implant Insertion Depth Estimation

Berk Takıt



SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

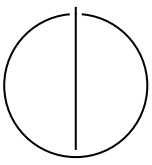
TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

AI-Based Cochlear Implant Insertion Depth Estimation

KI-Basierte Schätzung der Einführtiefe von Cochlea Implantaten

Author:	Berk Takıt
Examiners:	Prof. Dr. Wilhelm Wimmer, Prof. Dr. Björn Schuller
Supervisor:	Dr. Stephan Schraivogel
Submission Date:	10.03.2025



I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, 10.03.2025 Berk Takıt

Acknowledgments

This thesis marks the culmination of a challenging yet rewarding research journey and the conclusion of my studies at TUM. Throughout this time, I have been fortunate to receive the support and guidance of many individuals, and I would like to take this time to express my gratitude.

First and foremost, I sincerely thank Dr. Stephan Schraivogel for his unwavering support, invaluable insights, and mentorship. He always set aside time each week to guide and encourage me. A special mention goes to little Emilia, who unknowingly accompanied some of our calls—I hope one day she finds as much joy in learning as her father inspired in me.

I also extend my sincerest gratitude to Prof. Dr. Wilhelm Wimmer for his guidance, encouragement, and for allowing me to be a part of this project. His enthusiasm for his work is truly inspiring, and his insightful feedback from the very beginning has played a crucial role in shaping this thesis.

My thanks also go to Manuel Milling and Andreas Triantafyllopoulos for their valuable input and to all those involved in collecting and processing the data for this study.

The value of any pursuit is measured by the friends who stand beside us, and I am grateful to mine for their constant support. Special thanks to Lara, Can, and Baykam for helping me prepare for my midterm presentation—and to Batuhan, who tried his best, even if he ultimately slept through it. A big thank you to the rest of *Hamsiler*, whose friendship makes life all the more vibrant.

Above all, I am deeply grateful to my parents, Leyla and Ediz, for their unwavering love and encouragement. Their belief in me has been my guiding force. I also thank my sister, Buse, and hope I make her proud.

Lastly, a very special thank you to my one and only Bensu, whose belief in me never wavered. She has been a beacon of warmth for the last six years, and I look forward to cherishing many more by her side.

Fittingly, my research revolved around the ear—just like my mother's maiden name, Kulak, which means 'ear' in Turkish. Perhaps this journey was always meant to be.

Abstract

Cochlear implants (CIs) are neural prostheses designed to restore auditory perception in individuals with severe to profound sensorineural hearing loss. The precise localization of CI electrodes is crucial for hearing outcomes. The clinical standard for localization is post-operative computed tomography (CT) scans, but these expose patients to ionizing radiation and can increase cancer risk. Radiation-free alternatives for localization are thus desirable. This study aimed to develop multimodal approaches based on three input modalities - pre-operative CT scans, impedance measurements from the CI, and cochlear shape parameters - to develop deep learning models capable of predicting the linear insertion depth of the most basal electrode in the electrode array.

Different unimodal and multimodal models were created. Intermediate and late fusion approaches were investigated for each multimodal model. The performance of the models was then analyzed using a 10-fold cross-validation scheme, and compared to a mean model baseline. Our dataset consisted of 142 cases, all with the same implant manufacturer. After model selection, an ablation study with three different ablation methods was conducted to compare the contribution of each modality to model performance. Fully- vs. partially-inserted CI classification performance was analyzed based on a single training run on training and validation datasets with similar class distribution. Finally, the performance of the best-performing multimodal model was compared against state-of-the-art methods.

The best multimodal model, based on 3D-convolution layers and an intermediate fusion approach, achieved an absolute error of 0.98 mm $\pm$ 0.23 mm (mean $\pm$ standard deviation) compared to ground truth from CT scans. These results were comparable to state-of-the-art results. Ablation studies showed that each modality contributed to the model's performance. Impedance data seemed to be the most important modality, followed by CT scans, and lastly cochlear shape parameters.

The results from this study demonstrated that a multimodal approach to linear insertion depth regression was viable and showed promise for better performance than current models. The ablation studies and the results from intermediate vs. late fusion approaches suggested that joint feature representations learned using more than one modality were beneficial for this task. Future work should focus on improving multimodal models through better feature extractor architectures and more robust hyperparameter tuning.

Contents

1. Introduction

Along with written language, auditory communication is one of the pillars of human interaction, enabling individuals to engage in social, educational, and professional activities. The ability to perceive and process sound plays a crucial role in speech and language development, cognitive function, and even motor skills development [1]. However, hearing loss is one of the most prevalent neural impairments in the world. It is estimated that today around 1.5 billion people, including many children, are affected by some amount of hearing loss, with 400 million of those having moderate to complete hearing loss [2]. While mild hearing loss is not a major concern, moderate to complete hearing loss can lead to delayed language acquisition, social isolation, reduced educational and employment opportunities, and a decline in mental health if left untreated [2]–[4]. Furthermore, untreated hearing loss is estimated to cost over $980 billion annually in the form of related healthcare, education, productivity losses, and societal costs [2]. Perhaps even more alarming is that the number of people suffering from hearing loss is projected to considerably increase in the coming decades, growing to about 2.5 billion people worldwide by 2050 [2].

Hearing loss can arise from a variety of causes and is broadly categorized into three categories: conductive, sensorineural, and mixed [5]. Conductive hearing loss is the general term given to hearing loss caused by mechanical disruptions along the sound conduction pathway, including obstructions or damage in the external ear, the tympanic membrane, or middle ear ossicles, which prevent or reduce transmission of sound vibrations to the inner ear [6]. These disruptions can result from a variety of conditions, including obstructions, infections, or trauma affecting the external and middle ear [7]. Since the cochlea and auditory nerve are unaffected in cases of conductive hearing loss, it is often treatable through medical or surgical interventions and in cases with more serious damage to the external and middle ear structures, bone conduction hearing aid devices may be used to bypass the damaged structures and improve sound transmission to the inner ear [6].

In contrast, Sensorineural Hearing Loss (SNHL) refers to hearing impairment resulting from a combination of two dysfunctions: sensory hearing loss, caused by damage or degeneration of the hair cells within the cochlea, and neural hearing loss, which occurs due to dysfunction of the cochlear nerve, affecting the transmission of auditory signals to the brain [6]. SNHL has a diverse range of causes, which can be congenital

or acquired later in life. Congenital SNHL may stem from genetic mutations, infections during maternity (such as cytomegalovirus or rubella virus), prematurity, complications during birth (such as asphyxiation), and either dominant or recessive mendelian inheritance [6], [8]. Acquired SNHL can be the result of many causes such as exposure to ototoxins (such as some medications and industrial substances), viral infections (such as meningitis), head trauma, and, most commonly, trauma caused by exposure to loud noise (occupational, recreational, or accidental) [6]–[8]. More recently, exposure to COVID-19 has been suggested as a cause for sudden SNHL [9]. As the cause and severity of SNHL is very diverse from individual to individual, treatment options can also differ significantly. For individuals with severe to profound SNHL conventional hearing aids are often insufficient.

Cochlear Implants (CIs) are widely regarded to be an effective solution for such cases, particularly when damage to inner ear structures renders conventional hearing aids insufficient [10]. While both conventional hearing aids and CIs receive sound through microphones, CIs differ in that they do not simply amplify sound but instead convert it into electrical signals that directly stimulate the cochlear nerve, bypassing most of the auditory chain. The inner unit of the device, which is surgically implanted, directly stimulates the spiral ganglion cells of the auditory nerve via an electrode array inserted into the cochlea [11]. This electrical stimulation enables auditory perception by bypassing damaged inner ear structures. While CIs are the most common solution, other implants, such as auditory brainstem implants, may be used in specific cases. Precise localization of CI electrodes is important for the performance of CIs. The current clinical gold standard for the localization of CI electrodes is through the use of radiography, with the most widely used modality being Computed Tomography (CT) scans [12]. CT scans provide detailed and accurate images of the cochlea and electrodes, making precise localization possible. Use cases for localization include: intra-operative tracking, post-operative position verification, detection of extracochlear electrodes, monitoring of electrode migration, and individualized "fitting" based on tonotopy (i.e. the systematic arrangement of neurons based on their response to tones of different frequencies). The fitting process (also known as CI programming) involves the fine-tuning of parameters (such as pulse-width, rate of stimulation, and volume) so that the CI is optimized to generate electrical signals that yield the highest possible speech intelligibility for that specific individual [13], [14]. These adjustments are tailored to the individual's hearing thresholds and preferences, typically based on default 'Maps' in standard clinical practice. Although electrode positioning is generally not factored into the fitting process, emerging techniques like anatomy-based fitting aim to incorporate electrode locations for more personalized optimization [15].

Although localization is critical for the hearing outcome, two major drawbacks of using CT scans make it worthwhile to explore alternative options. First, CT scans

expose patients to doses of ionizing radiation which in turn can increase the risk of cancer. Studies have shown a positive correlation between CT scans and cancer risk, especially for children [16]–[19]. Second, CT scans can be costly, both in terms of the equipment required and in terms of the specialized personnel to operate the machines and interpret the results. This makes CT less accessible, particularly in settings with limited resources, increasing the overall cost of CI surgeries and follow-up care. These two drawbacks combined make an alternative that reduces both radiation exposure and costs while still ensuring accurate electrode localization very desirable.

Deep Learning (DL), a branch of Machine Learning (ML) and Artificial Intelligence (AI) that utilizes multi-layered neural networks, has matured significantly over the past decade. DL models excel at identifying complex, non-linear relationships in large and diverse datasets - a capability that has made them highly successful in healthcare applications [20]. Building on these strengths, recent research has applied DL to leverage multiple modalities (such as Magnetic Resonance Imaging and CT scans) as inputs [21]. The integration of multiple modalities is supposed to give DL models a more comprehensive understanding of complex medical information, ultimately leading to better performance [22]. In the context of CIs, DL holds considerable potential for improving electrode localization performance while reducing dependency on post-operative radiographic imaging by leveraging an input combination of pre-operative CT scans, intra-operative Impedance Field Telemetry (IFT) recordings, and patient-specific information.

1.1. Background and Motivation

To address the shortcomings of CT scans as the golden standard for electrode localization, alternate approaches for localization using the IFT recordings from the CIs have been suggested [23]–[30]. These approaches rely on a combination IFT recordings made during or after the operation, cochlear dimensions measured from pre-operative CT scans, and demographic data for electrode localization. Some of these methods demonstrate promising results in predicting the Linear Insertion Depth (LID) (the linear distance in millimeters between the round window of the cochlea and the most basal electrode in the electrode array) and aim to significantly reduce reliance on radiographic imaging [23], [25], [29], [30]. As a next step, it may be possible to combine pre-operative CT scans with IFT data to further improve performance by leveraging the non-linear relationships between these two modalities and the LID. DL is particularly well-suited for this task, as it excels at capturing complex, non-linear relationships that may not be immediately apparent through classical ML (linear regression, support vector machines, decision trees, etc.) or human interpretation. By leveraging DL methods, we may be

able to exploit patterns and dependencies between these two modalities that may be difficult to model explicitly, ultimately leading to more precise electrode localization while still reducing the need for post-operative CT scans.

1.2. Problem Statement

Given a dataset containing multiple modalities — including pre-operative CT scans, intra-operative IFT recordings, cochlear shape parameters, and demographic data — this project aimed to develop DL models that used these inputs with the primary task of predicting LID values. Additionally, estimating LID enabled the secondary task of classifying CI insertions as either fully-inserted (all electrodes inside the cochlea) or partially-inserted (some electrodes remaining extracochlear). These tasks are clinically relevant for intra-operative assistance, post-operative assessment, and treatment planning.

1.3. AITIDE Project

The AITIDE Project builds upon the findings of the ITIDE study, aiming to refine impedance-based estimation of CI LID through the use of ML. While the ITIDE project demonstrated that impedance and field telemetry could be used for radiation-free electrode localization, its initial phenomenological model struggled with accuracy, particularly for partially-inserted electrode arrays. In response, AITIDE seeks to develop and optimize ML approaches that outperform traditional statistical models in both sensitivity and specificity. The project aims to produce an industry-ready algorithm capable of providing reliable postoperative electrode localization without the need for radiographic imaging. As part of the AITIDE project, this thesis explores the possibility of using pre-operative CT scans together with IFT data to build a DL model capable of estimating LID.

1.4. Research Objectives

The primary objective of this project was to develop a Deep Learning (DL) model that utilized multiple modalities to regress Linear Insertion Depth (LID) values of the most basal electrode in the Cochlear Implant (CI). This involved:

1. Exploring different architectures for multimodal DL models.

2. Evaluating model performances using quantitative metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) with cross-validation to select the best-performing model.

3. Comparing the proposed model's performance with baseline models and existing state-of-the-art results to determine whether incorporating multiple modalities, such as pre-operative CT scans and IFT data, as input provided an increase in performance.

The secondary objective was to evaluate the performance of the best-performing model from the first objective on the binary classification task of fully- or partially-inserted CIs.

1.5. Scope of the Study

The focus of this study was the development, implementation, and evaluation of multimodal DL models for CI LID estimation. The dataset used, consisting of 142 cases, originated from multiple medical centers and was collected as part of the AITIDE project. No new data collection was conducted as part of this study. Existing data and methodologies from previous AITIDE studies were used for ground truth extraction.

2. Related Work

2.1. Cochlear Implants

Cochlear Implants (CIs) are neural prostheses designed to restore auditory perception in individuals with severe to profound Sensorineural Hearing Loss (SNHL). Unlike conventional hearing aids, which amplify sound, CIs bypass the damaged sensory hair cells in the cochlea and directly stimulate the auditory nerve using electrical signals. This allows individuals with significant hearing impairment to perceive sounds and, in many cases, develop speech recognition capabilities. The use of CIs represents the first substantial restoration of a human sense, a significant and groundbreaking achievement in biomedical engineering [31].

While the concept of electrical stimulation of the auditory nerve goes farther back, the invention of the first viable CI dates back to the 1960s. In the early 1960s, Dr. William Fouts House, widely recognized as the "Father of Neurotology" and credited with inventing the first single-channel CI, began working on CIs. However, his early designs encountered significant biocompatibility issues, which limited their effectiveness [32]. Due to these issues, Dr. House postponed any further work on CIs until 1967, when advances in biocompatible materials allowed his work to continue along with electrical engineer Jack Bauer, and together they produced and implanted into several patients a viable CI system that could for the first time be reliably used outside of a lab environment [31], [32]. Dr. House's work culminated in the development of the "3M House Cochlear Implant System", which became the first CI to receive approval by the US Food and Drug Administration (FDA) for use in patients; which not only validated decades of research into CIs, but also marked the first time a medical device was officially recognized for restoring a human sense [32], [33].

Dr. House favored single-channel implants over multichannel implants, citing the simpler design which in turn allowed prices to be lower and thus more accessible [32]. However, this approach limited speech comprehension as complex auditory signals could not be effectively conveyed. Recognizing this need for greater sound resolution, multiple research teams around the world worked on developing multi-channel implants; including the team led by Graeme Clark in Australia and the team led by Ingeborg Hochmair in Austria [31]. These efforts ultimately led to the first multichannel implants, which allowed for stimulation at different points along the

cochlea, significantly improving speech perception by replicating the cochlea's natural tonotopic organization [31], [32]. This improvement over single-channel meant that multichannel implants became the standard, which is still the case today with ongoing advancements in signal processing, electrode design, and minimally invasive surgical techniques for implantation [31]. As of 2022, it is estimated that over a million CIs have been implanted in individuals worldwide, making them the most widely used type of neural prosthesis [34]. Currently, the global cochlear implant market is dominated by three major manufacturers: Cochlear Limited (Sydney, Australia), MED-EL (Innsbruck, Austria), and Advanced Bionics (Valencia, CA, USA) [35].

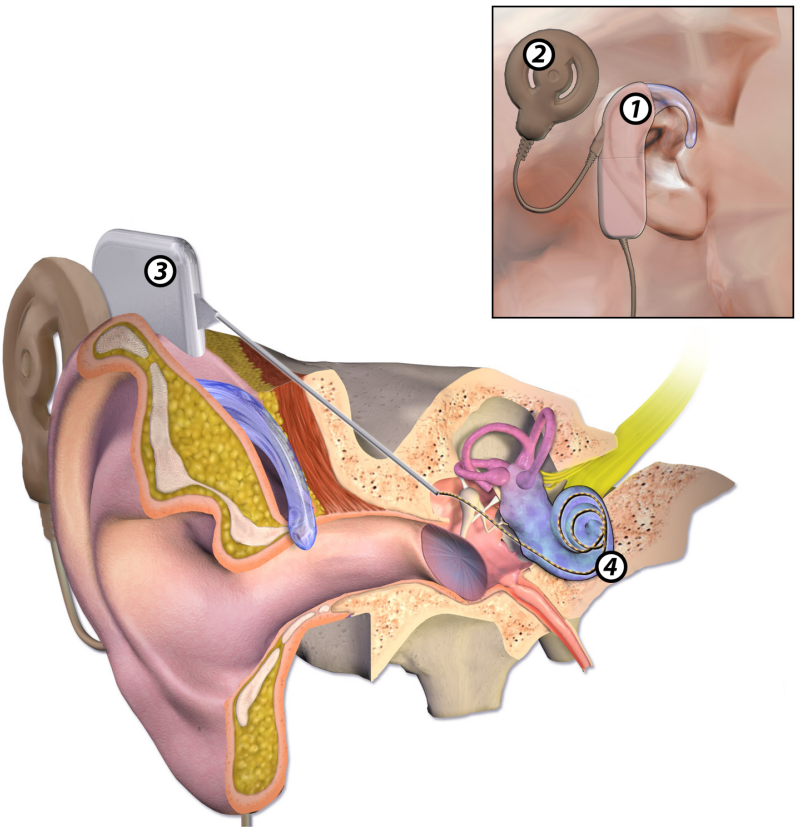


Figure 2.1.: Diagram showing the external and internal components of a cochlear implant. The external part consists of a unit containing the microphones and the audio processor (1), and a transmitter coil (2). The internal components are the receiver coil (3), and the electrode array inserted into the cochlea (4). *Adapted from [36], under the Creative Commons License CC BY-SA 4.0.*

A cochlear implant consists of two main components (Figure 2.1): an external part

that sits behind the ear and an internal part that is surgically implanted under the skin. The external part includes microphones to pick up sound from the environment, a speech processor to convert these sounds into electrical signals, and a transmitter coil that sends signals to the internal unit via radio frequency. The internal part consists of a receiver coil to receive the information sent by the external unit that is then converted into electric impulses and sent to the second component, an electrode array, typically consisting of between 12 and 24 electrodes, that is inserted into the cochlea to directly stimulate the auditory nerve.

2.2. Impedance Telemetry

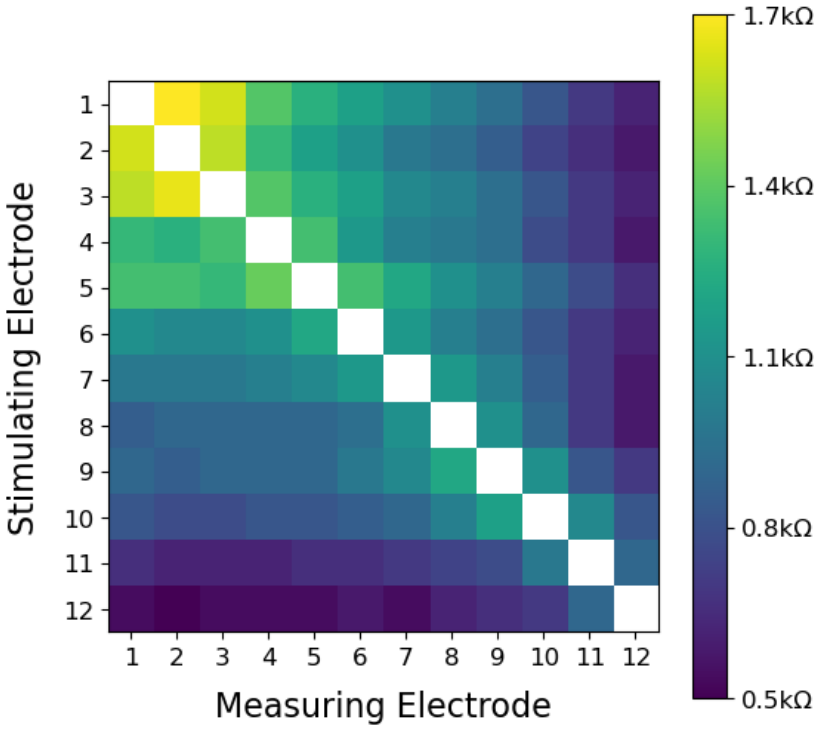


Figure 2.2.: Heat map visualization of a 12x12 impedance matrix. Note that the diagonal values (which represent cases where the stimulating and receiving electrodes are the same) are masked to avoid distorting the color scaling as they are significantly larger than the off-diagonal values.

Impedance telemetry in CIs is a diagnostic technique used to assess the electrical properties of the implanted electrode array and its interaction with the surrounding tis-

sue. It is done by delivering a small, controlled electrical current through a stimulating electrode in the array, while another electrode (the receiving electrode) measures the resulting voltage response with respect to a reference electrode that is most commonly located on the housing of the internal unit of the CI. The ratio of the measured voltage to the applied current allows for the calculation of impedance values, which can be represented as an $N \times N$ matrix, with N being the number of electrodes in the CI electrode array. Each value in this matrix represents the impedance measured at a receiving electrode in response to the applied current at a stimulating electrode, with respect to a reference electrode (Figure 2.2).

Clinically, impedance telemetry is routinely used to verify electrode integrity and detect critical issues such as short circuits or open contacts [13]. In addition to its role in device diagnostics, and given that it can be recorded both during operation and post-implantation, impedance telemetry offers a practical and radiation-free alternative for extracochlear electrode detection, electrode migration monitoring, and electrode localization.

2.3. Computed Tomography Data

Computed Tomography (CT) scans are a widely used medical imaging technique that produces detailed cross-sectional images of the body by utilizing X-rays and computer processing. CT scans were first used in a clinical setting in 1972 and revolutionized medical imaging by becoming the first slice-imaging modality, allowing a volumetric view of the body's internal structures [37]. Unlike conventional radiography that only generates 2-dimensional projections, CT scans allow for a 3-dimensional reconstruction of the scanned area of the body, making it very valuable for a wide variety of applications including diagnostics, monitoring, and surgical applications. One of the more common CT technologies is cone-beam CT, which delivers high-resolution scans while exposing patients to lower doses of radiation [38]. Cone-beam CT scans were initially used almost exclusively for dental radiology, but have since been used in a much wider variety of non-dental applications, including CI electrode localization [39]. The most recent advancements in CT technology are photon-counting detector CT scanners that can deliver better resolution images than other methods [40]

CT scans play a crucial role in the pre-operative, intra-operative, and post-operative phases of CI implantation. In the pre-operative phase, CT scans are essential for candidate assessment and surgical planning [41], [42]. CT scans provide a detailed volumetric visualization of the cochlear anatomy, enabling clinicians to identify any malformations, ossification, and anatomical variations that can influence the implantation procedure. This information can then be used for: determining whether the candidate is suitable

for a CI, choosing an appropriate electrode array, and planning the surgical trajectory. Intra-operative CT scans are used to confirm proper positioning of the CI, allowing for repositioning with minimal increase in operation time [12]. In the post-operative phase CT scans are used for confirming the positions of the electrodes inside the cochlea and for monitoring electrode migration [12], [43], [44]. The verification of electrode positioning is of particular importance, as misplacement of electrodes (such as extra-cochlear electrodes) can adversely affect auditory outcomes. Monitoring the changes in electrode locations over time aids in optimizing CI programming and guiding potential corrective interventions.

CT scan data is typically stored in the Digital Imaging and Communications in Medicine (DICOM) format, the standard for biomedical imaging [45]. DICOM files contain both the image data and associated metadata, including acquisition parameters, patient metadata, and scanner settings. DICOM files usually contain a single slice of the volumetric scan, so one CT scan may generate hundreds of individual DICOM files that are stored sequentially. For more efficient storage and processing, CT scan data is often converted from DICOM format to Nueuroimaging Informatics Technology Initiative (NIfTI) format, especially for data analysis and other computational applications [46]. NIfTI files offer a more streamlined structure by storing the CT scan in a 3-dimensional format in a single file.

A crucial aspect of CT imaging is the Hounsfield Unit scale, which quantifies substance density [48]. Different kinds of substances exhibit distinct values, allowing for differentiation between bone, soft tissue, air, fluids, and foreign bodies. The 3-dimensional pixels, or voxels, that compose a CT scan typically store intensity values in Hounsfield Units. Another crucial aspect to take into consideration when working with CT scan data is the voxel dimensions, usually stored as two fields in the metadata. The height and width of a voxel, which is the physical distance between the centers of adjacent pixels in the acquisition plane, are given by the pixel spacing field; and the depth of a voxel, which is the physical distance between two consecutive slices in a scan, is given by the slice thickness field. CT scans can be acquired from any arbitrary orientation, but most often the acquisition plane is one of the three principal anatomical planes: axial, sagittal, or coronal (Figure 2.3).

2.4. Electrode Localization

Electrode localization refers to the process of determining the precise location of CI electrodes with respect to anatomical landmarks. Accurate localization is crucial for both post-operative position verification and CI programming tasks. The clinical standard for electrode localization is using post-operative CT scans. The annotation

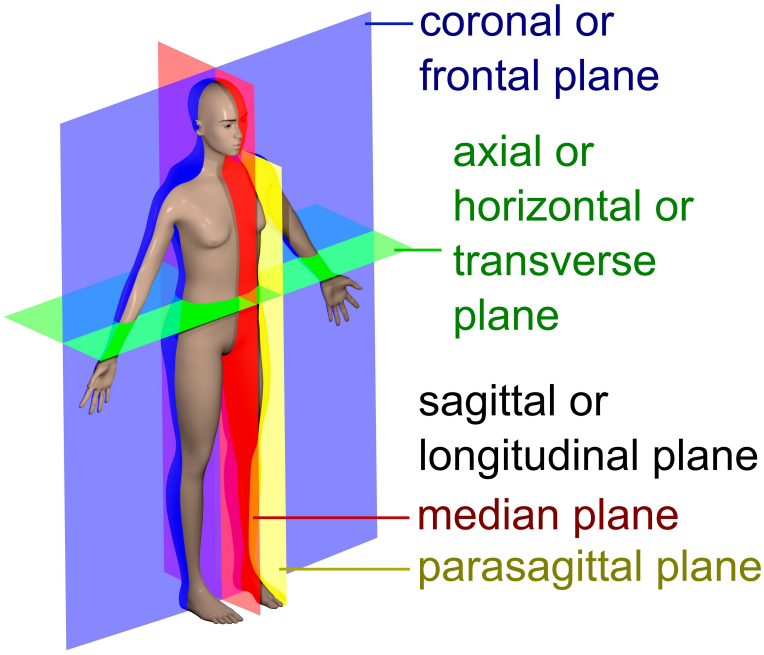


Figure 2.3.: Illustration of human anatomical planes. *Used as is from [47], under the Creative Commons License CC BY-SA 4.0.*

process is typically done manually by trained professionals, necessitating the careful inspection of the CT scans to locate and label each electrode. Alternatively, methods using IFT data recorded from CIs have also been shown to be a viable way for electrode localization without the need for manual labeling [24]. This is made possible by the fact that impedance measurements are affected by the surrounding environment of the electrode [25].

2.5. AI in Medical Applications

Artificial Intelligence (AI) has become a mainstream tool for use in modern medical applications such as anomaly detection from various imaging modalities, disease diagnostics, personalized treatment, drug development, and even physical tasks such as robotic-assisted procedures [49], [50]. One of the more significant areas of medicine where AI, and more specifically DL, has demonstrated remarkable progress and results is medical imaging. DL models have been applied extensively to imaging tasks, including tasks involving CT scans [51]. The tasks for these models include automated

segmentation, classification, regression, and registration. Automated segmentation has particularly gained traction in medical imaging, with models based on Convolutional Neural Network (CNN) architectures [52], performing segmentation of anatomical structures, tumors, lesions, and many more on different kinds of medical imaging modalities [53]. Models trained for classification tasks have also shown state-of-the-art performance using similar architectures to segmentation to distinguish between different classes such as normal and pathological tissues [54]. These AI-driven approaches help to reduce the manual workload and variability in clinical practice while also enhancing the accuracy of medical assessments, with performance levels often reaching or even exceeding professionals [53].

AI has also been applied to the field of Otology, where DL techniques have been used for a variety of tasks. One of the primary applications has been cochlear segmentation, where DL models have been utilized to automatically segment the cochlea from CT scans, with precision close to manual segmentation [55]. DL has also been used for localization of the cochlea from CT scans, automatically detecting the cochlea and extracting cochlear measurements (such as cochlear duct length and basal diameter) [56]. These applications can be used in the pre-operative phase of CI implantation for guiding the surgical planning process. In the post-operative phase, DL has been used for the automatic detection and localization of CI electrodes using post-operative CT scans as input [57], [58]. These approaches can identify the position and orientation of electrodes within the cochlear duct without human intervention, decreasing the workload for professionals. However, these methods still rely on post-operative CT scans as input, necessitating further exploration of alternative methods for estimating electrode positions to reduce the radiation exposure of patients post-operation.

Despite the significant advancements and successes of DL in medical applications, several challenges are still present that hinder efforts for further clinical adoption. One of the most pressing issues is the limited availability of high-quality, annotated datasets. Medical imaging datasets are often smaller in size due to several limitations in collecting data such as patient privacy, ethical regulations, and limited data for rare conditions [59]. This makes DL with CT scans and other such modalities very difficult as the number of parameters needed to learn meaningful representations of these complicated inputs usually vastly outnumbers the amount of training data available. Another critical challenge is data variability and standardization. Medical images can vary significantly based on differences in scanning protocols, imaging equipment, and artifacts introduced during the imaging process; making it difficult for DL models to generalize effectively. These challenges require the pre-processing of imaging data to ready them for model training.

2.6. Multimodal Deep Learning

Multimodal deep learning is the concept of integrating multiple sources of data to enhance a model's predictive capabilities by leveraging complementary information from different modalities such as images, video, text, and audio. While the idea of multimodal learning is not new, the concept was first applied to DL in the early 2010s as researchers sought to improve feature learning using multiple modalities [60]. This approach to DL is especially approrpriate for medical applications as this domain features a wide variety of input modalities that can be used for a task; ranging from imaging modalities such as CT scans and various types of Magnetic Resonance Imaging (MRI) modalities to omics data, demographic data, and more [22]. Leveraging multimodal learning in this context enables models to combine features extracted from each of these separate modalities and thus perform better than models trained on a single modality. Additionally, multimodal models address challenges such as missing data and noise by utilizing cross-modal information, enhancing robustness and reliability. This means that if one type of data is unreliable or unavailable, the model can rely on the remaining sources of data to maintain performance. Multimodal deep learning has been applied to many tasks in medical applications, such as brain tumor segmentation by combining different types of MRI, classification of breast cancer using MRI images and clinical features, and COVID-19 severity classification using CT scans and laboratory indicators to name a few [21], [61], [62].

The fusion of different modalities is a key concept in multimodal learning. Fusion refers to the process of integrating information from different modalities to create a more comprehensive and effective overall representation [64]. This can be done at various levels; namely feature-level or early fusion, decision-level or late fusion, and layer-level or intermediate fusion (Figure 2.4) [22]. The choice of fusion approach can be based on the correlation of modalities, modality-specific limitations, task-specific considerations, data availability, and computational efficiency.

2.6.1. Early Fusion

Early or data-level fusion integrates multiple sources of data by concatenating them into a single input vector before feeding it into the model [63]. Each modality can be added to the input vector as raw data, or as either handcrafted or learned higher-level representations [61]. This can lead to large input feature vectors that may contain redundant information, but this issue can be overcome with dimensionality reduction techniques such as principal component analysis [61]. The difference between early fusion and intermediate fusion when using feature extraction is that in intermediate fusion the loss is propagated back to the feature extractor for training, whereas in early

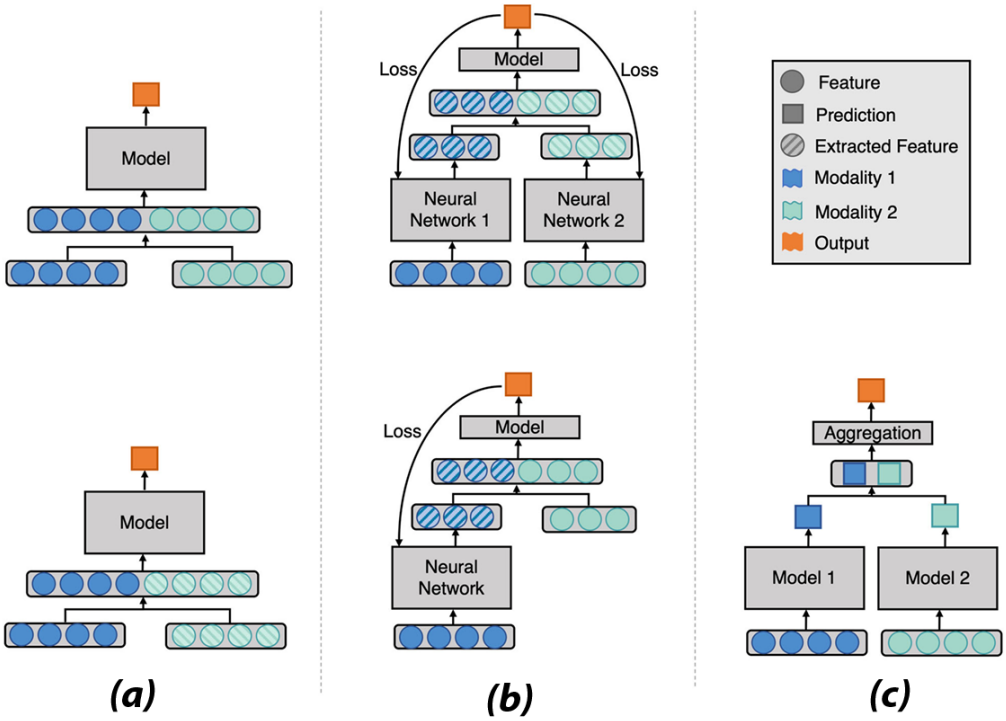


Figure 2.4.: Overview of different fusion methods. (a) Early or data-level fusion, (b) intermediate or layer-level fusion (also referred to as joint fusion in the original figure), (c) late or decision-level fusion. *Adapted from [63], under the Creative Commons License CC BY 4.0.*

fusion the feature extractor is not trained [63].

2.6.2. Intermediate Fusion

Intermediate or layer-level fusion combines information from different modalities at deeper layers within the model. Compared to early fusion, intermediate fusion allows models to extract modality-specific features before integration, reducing redundant or noisy information that might arise from raw feature concatenation [61]. Similarly, unlike late fusion which only merges high-level decisions, intermediate fusion facilitates deeper interactions between modalities, leading to richer feature representations and improved generalization across tasks [61]. In this approach, modalities are processed separately by a shared model and mapped to higher-level representations before ultimately being fused into a shared representation. The fusion process can be incremental, where

representations are integrated at multiple levels throughout the model or performed all at once [63]. This shared representation is then fed to a final network which typically consists of a series of fully connected layers followed by activation layers to produce a final output. This approach is particularly suitable for medical applications as it enables the combination of spatial information extracted from imaging data with structured patient data [21].

2.6.3. Late Fusion

Late or decision-level fusion uses predictions from multiple sub-models, each trained on a single modality, that are aggregated to make a final prediction [64]. This approach is particularly useful in cases where some modalities may be missing for some cases or when modalities are not very correlated [61]. Ensemble learning techniques such as majority voting and weighted averaging can be employed to aggregate predictions from separate models.

2.7. State-of-the-Art

Approaches to electrode localization using IFT data were shown to produce clinically viable results by several studies. *Bruns et al.* [24] used a nonlinear model and a Long Short-Term Memory (LSTM) model with impedance data to produce cochlear region predictions. Similarly, *Dong et al.* [26] used impedance data and access resistance for the detection of electrode translocation. Additionally, *Sijgers et al.* [28] and *Giardina et al.* [27] both explored the relationship between impedance changes and electrode positioning, reinforcing the viability of using impedance data for electrode localization.

Insertion depth estimation has been an active area of research. *Zhang et al.* [30] used transimpedance matrices and a non-linear model to predict angular insertion depths. *Aebischer et al.* [25] demonstrated an impedance-based method for intra-operative estimation of LID, with a MAE of 0.76 ± 0.53 mm. Their approach utilized bivariate spline extrapolation to estimate tissue resistances from transimpedance recordings, which combined with a phenomenological model was able to estimate LIDs. Building upon this, *Schraivogel et al.* [29] extended this approach to post-operative settings, where their phenomenological model was able to estimate LIDs with a MAE of 0.9 ± 0.6 mm, further solidifying the feasibility of impedance-based LID estimation. More recently, *Schraivogel et al.* [23] employed machine learning models trained on impedance data, features extracted from the impedance data, cochlear shape parameters, and demographic data to estimate LIDs. Their model based on Extremely Random Trees, or Extra Trees for short, achieved a MAE of 0.8 ± 0.6 mm (less than half the interelectrode

distance of the arrays used in their study, which was 2.1 mm), outperforming the phenomenological models; especially in cases with extracochlear electrodes.

3. Methodology

3.1. Dataset

The dataset used in this study was collected as part of the AITIDE project and comprises 142 cases. Of these cases, 116 were already annotated and 26 were annotated as part of this study. All CI electrode arrays in the dataset were produced by the same manufacturer (MED-EL, Innsbruck, Austria) and featured 12 electrodes. The data used from each individual case for this study was as follows:

- **Pre- and post-operative CT scans** in DICOM format.
- **Impedance matrix** based on IFT data recorded from the CI.
- **Cochlear landmarks** and **electrode positions** manually annotated from the CT scans.
- **Cochlear shape parameters** derived from the annotated landmarks.
- **LID values** and **number of extracochlear electrodes** obtained using the annotated landmarks and electrode positions.

The pre-operative CT scans, impedance matrices, and cochlear shape parameters were used as input for the models. The cochlear landmarks were used during the preprocessing steps and the LID values and whether there were any extracochlear electrodes were used as target values during training and validation for the regression and classification tasks, respectively. An overview of all cases in the dataset can be seen in Table C.1.

3.1.1. Impedance Matrices

The impedance matrix for a case was generated from the IFT data recorded from the CI after implantation. The values (denoted by Z, in Ω) were obtained by dividing the recorded voltage (in V) by the stimulation current (in A) for each pair of stimulating and receiving electrodes. As each electrode array in the dataset consisted of 12 electrodes, this yielded a 12×12 matrix of impedance values. These values were then standardized

with z-score standardization using the formula $z = \frac{x-\mu}{\sigma}$ where z is the standardized and unitless value (also called the z-score or standard score), x is the original value, μ is the mean, and σ is the standard deviation of the values in the matrix.

3.1.2. Annotation of CT Scans

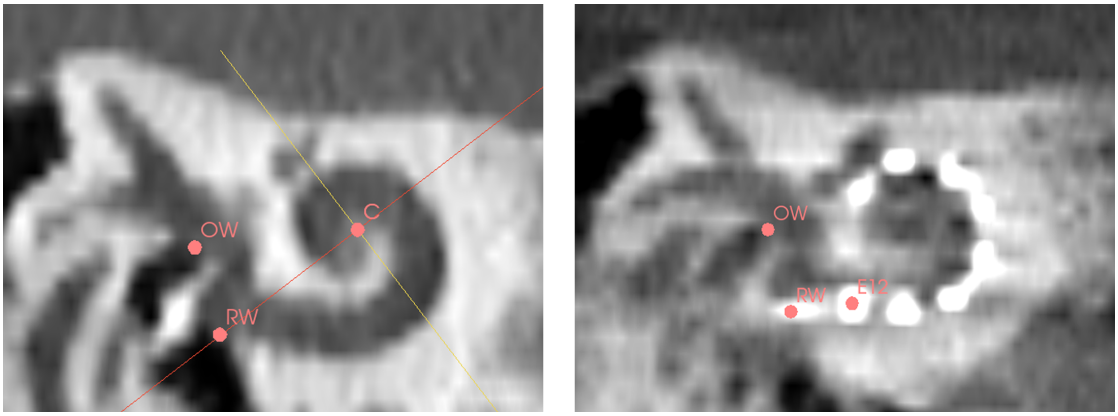


Figure 3.1.: Two images showing slices from the same case, from a pre-operative (left) and from a post-operative (right) CT scan, annotated using the process described in Section 3.1.2. The slices have been re-oriented to show the cochlea in the so-called cochlear base view. The pre-operative image shows the center of the cochlea (C), the round window (RW), and the oval window (OW). The post-operative image shows the most basal electrode (E12) as well as the OW and RW landmarks.

The annotation of CT scans was done manually using 3D Slicer, a free and open-source software for the visualization and processing of 3D images [65]. The annotation steps were as follows (Figure 3.1 shows the state at two points in the annotation process):

1. The pre- and post-operative CT scans were cropped around the cochlea and then the post-operative scan was registered onto the pre-operative scan.

2. The slices were repositioned (translated and rotated) to display the cochlear cross-section fully.

3. Cochlear landmarks were marked on the pre-operative scan and their coordinates were extracted. These landmarks were the middle of the round window (RW), the center of the cochlea (C), the apex of the cochlea (A), and the middle of the

oval window (OW). Of these landmarks, the most important ones for this study were RW and C; as RW was used for ground truth extraction, and C was used in the CT data preprocessing pipeline.

4. The cochlear base length (a value), cochlear base width (b value), and cochlear height (h value) were measured on the pre-operative scan using the landmarks annotated in the previous step [66].

5. The 12 electrodes were marked on the post-operative scan as E1 through E12, with E12 being the most basal and E1 being the most apical electrode, and their positions were extracted.

The coordinates of the landmarks and the electrodes are based on the local coordinate system of the CT scan, where the C landmark is at the origin $(0, 0, 0)$. The LID of each electrode was then extracted by calculating the linear distance from the RW landmark to the electrode in millimeters. The ground truth for this task was the LID of the electrode labeled E12, the most basal electrode in the array. Then the number of extracochlear electrodes was calculated by counting the number of electrodes with negative LIDs. The data was then saved in JavaScript Object Notation (JSON) format for use in model training and validation.

3.1.3. Preprocessing of CT Scans

The CT scan data was represented by a 3D array of values, with each value representing the intensity of a voxel in the image in Hounsfield Units. There were many variations among CT scans that had to be standardized before they were ready to be fed to the models. To ensure consistency across cases and improve model generalization, preprocessing steps were applied to the CT scans. Some of these steps used custom-made functions while others used the MONAI package, a PyTorch-based framework for DL in healthcare [67].

The first step was to crop the CT scan around the center of the cochlea (cochlear landmark C) to get rid of unnecessary anatomical structures and center the image on the cochlea. To do this, the coordinates of the landmark were transformed from real-world coordinates to array indices, and then a rectangular prism was cropped from the array around the landmark point, ensuring an appropriate margin was maintained to preserve anatomical context and relevant structures. This step had the bonus of greatly reducing the image size and thus reducing computational complexity. The code for this step can be seen in Listing B.1.

The next step was to standardize the orientation to a common coordinate system. This was important as of the two most common bases, LPS (Left, Posterior, Superior)

and RAS (Right, Anterior, Superior), DICOM images use the former while 3D Slicer uses the latter [68]. The transformation from one system to the other was trivial and could be done by flipping the signs of the first two axes.

After this, the next step in the preprocessing pipeline was scaling the intensity of the scans, which was done by clipping values outside a given range and then scaling all values to a target range. This transformation ensured that the substances we wanted (soft tissue, bones) were emphasized while de-emphasizing and clipping substances that were less important for the task (air, fluids, foreign bodies such as metals). All intensities were then rescaled into a normalized range, making them more suitable for DL.

The most important step in the pipeline followed, where the scans were resampled to ensure uniform voxel dimensions across the dataset. This was crucial as when two scans have different voxel dimensions, two identical physical structures can have a different number of voxels, leading to different representations as the model does not know the specific voxel dimensions for each scan. By resampling all CT scans to the same voxel dimensions, we could ensure that spatial relationships and anatomical proportions remained consistent across cases, enhancing the robustness of the model by enabling it to extract more accurate features from the scans.

As the final step, all scans were resized to the same target shape using interpolation. This was done to ensure each scan matched the expected input shape for the models. The target shape varied based on the model used.

3.2. Model Architectures

The primary goal of all models in this study regardless of architecture was the regression of the LID of the most basal electrode. The secondary goal, which could be done by looking at predicted LIDs, was the classification of partially- and fully-inserted cases. The models took 3D arrays representing pre-operative CT scans, 2D arrays representing impedance matrices, and two scalars of cochlear shape parameters (namely a and b values, or cochlear base length and width) as input. While the shapes of the impedance matrix and cochlear shape parameters were consistent, 12×12 2D array and vector consisting of 2 features respectively, the shape of CT scans were different on a model-by-model basis. Models that took only one of the three modalities as input were created to analyze the performance of unimodal approaches. Different architectures were considered for each modality. Multimodal models taking all three modalities as input were created. These were grouped into two categories: models trained from scratch and models using other pre-trained models for CT feature extraction.

3.2.1. Single Modality Baseline Models

Cochlear Shape Models

As the shape parameter was a vector with only two values, these models consisted of the simplest neural networks: single- and multi-layer perceptrons (Figure 3.2). The difference between models was the number of hidden layers and the number of nodes in each layer. We tested five different architectures for this modality, each with an increasing number of hidden layers.

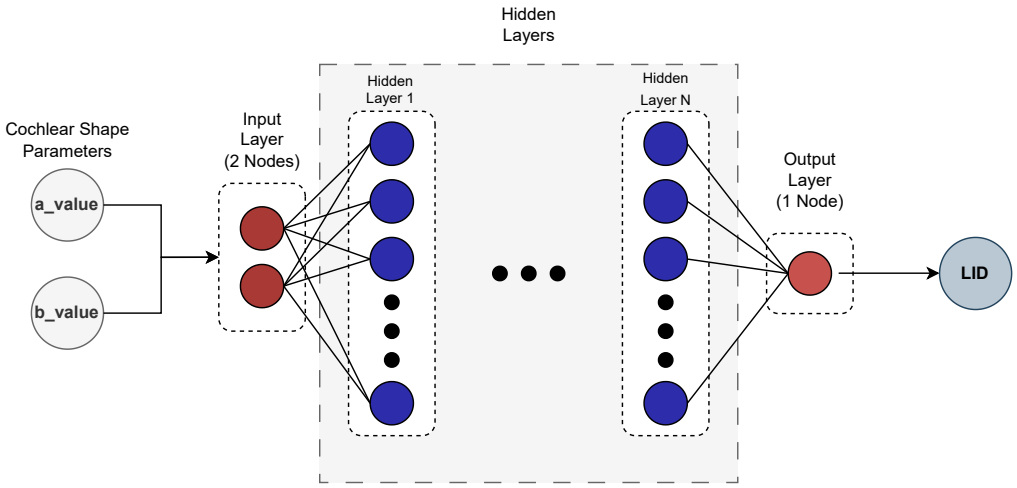


Figure 3.2.: Diagram for the general architecture of models taking only the cochlear shape parameters as input. Note that the number of hidden layers and the number of nodes in these layers were configurable, leading to different model architectures.

Impedance Matrix Models

For these models, two approaches were explored. First, we flattened the 12×12 arrays into vectors of 144 values, enabling the creation of single- and multi-layer perceptions capable of processing this modality. Second, we used 2D-convolution layers to process the matrices as spatial data, preserving local relationships that are lost when flattening the arrays into vectors. Unlike perceptrons, which treat each value independently, convolution layers can capture local patterns and relationships by applying learnable filters that scan across the data, extracting features from areas instead of relying solely on individual data point values.

For the first approach, five different models were tested, each with a different number of hidden layers and nodes. For the second approach, we tested four different CNN architectures, each consisting of the same convolution-activation blocks but with different depths and number of filters. Figure 3.3 shows the general architecture for these two approaches. The implementation of the first approach can be seen in listing B.2.

CT Models

Models using only the CT scans were a 3D-CNN based on 3D-convolution, max-pool, and activation layers; and a model using the Computed Tomography-Foundation Model (CT-FM) feature extractor [69] (Figure 3.4). Both models extracted the features from the CT scans before feeding them into a fully connected neural network to create the final output.

CNNs are well suited for this task as they can extract features from images and volumes at varying levels of complexity, constructing a rich representation while preserving spatial information in the data. However, training a CNN from scratch is most effective with a large dataset, as a substantial amount of data is required to learn filters that extract the most relevant features. As this was not the case for our study, we also included a model using a feature extractor that was trained on a much larger dataset compared to our own. The drawback of this approach was that the features extracted by the pre-trained model had a chance of not being transferable to our task.

The model using a 3D-CNN for feature extraction was trained wholly from scratch. In contrast, the model using the pre-trained CT-FM feature extractor only had its decision head trained (i.e., the fully connected neural network that was fed the features extracted by the feature extractor). The CT-FM feature extractor was frozen throughout the entire training run.

3.2.2. Multimodal Models

The performance of these models was the main focus of this study. They took all three modalities as input. The models were grouped into two broad categories based on how they extracted features from the CT scans: using 3D-CNNs trained from scratch or by using pre-trained models trained on different datasets, also known as transfer learning. The architectural diagrams of these models can be seen in Appendix A.

Models Trained from Scratch

In the first approach, 3D-CNNs were trained from scratch on our dataset. This method allowed for highly task-specific feature extraction but was very dependent on having

large amounts of data, which was not the case in this study. Three main architectures were explored for this approach: two architectures with different depths composed of 3D-convolution, pooling, activation, and dense layers (similar to the well-known LeNet architecture [70]), and an architecture based on the GoogLeNet architecture [71]. The shallower and deeper LeNet-like models were called 'Shallow CNN' and 'Deep CNN' respectively, and the model based on the GoogLeNet architecture was called 'GoogLeNet3D'.

Transfer Learning

In the second approach, transfer learning was utilized to leverage pre-trained models that were trained on much larger datasets than our own. Two different pre-trained models were used for feature extraction: the Inception-v3 model [72], and the CT-FM feature extractor [69]. These multimodal models were called the 'Inception-v3 Transfer' model and the 'CT-FM Transfer' model respectively.

Inception-v3 Transfer Model The Inception-v3 model was trained on the classification task of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [73], which consists of over a million RGB images. At the time of its publication, the model achieved state-of-the-art results. This model was chosen because of the inception modules that can capture multi-scale spatial features efficiently, making it well-suited for extracting representations from complex image data at multiple scales [72]. We believed this ability would transfer well to our task, as CT scans exhibit structural patterns at multiple scales, and leveraging a model capable of capturing these patterns could improve the overall model performance. Furthermore, initial testing showed that Inception-v3 performed better for our task compared to other models trained on the ILSVRC dataset (such as ResNet or EfficientNet). As the Inception-v3 model was trained for processing 2D RGB images, the slices of the pre-operative CT scan were first repeated into three channels to mimic RGB images and were then fed through one by one to extract feature vectors for each slice. Then, these feature vectors were stacked on top of each other before being fed into a bidirectional LSTM [74] module, which captured spatial dependencies across slice representations. The bidirectional nature of the LSTM allowed it to process sequences both forward and backward, ensuring the contextual information from adjacent slices was incorporated. This sequential representation was then passed through a simple attention mechanism that assigned importance weights to each slice and computed a weighted sum of the LSTM outputs. This process resulted in a single feature vector that represented the CT scan, which was then concatenated with the flattened impedance matrix and the cochlear shape parameters to be fed into the final fully connected neural network. The implementation

of this model can be seen in Listing B.3.

CT-FM Transfer Model The CT-FM was trained on over 140.000 CT images for radiological interpretation tasks such as whole-body segmentation and head CT triage classification; achieving superior performance compared to state-of-the-art models [69]. As these tasks are heavily dependent on extracting good higher-level representations of CT scans, we included it in our study to see if the extracted features translated well to our task. The 'CT-FM Transfer Model' used the feature extractor part of the pre-trained CT-FM model to extract features from pre-operative CT scans. Each feature was then standardized individually using pre-computed means and standard deviations specific to that feature. These statistics were obtained by processing our dataset through the feature extractor and computing the mean and standard deviation for each extracted feature. After this, features extracted from the impedance matrix using a fully connected neural network were concatenated with the standardized CT features and the cochlear shape parameters to create a single feature vector, which was then fed into the final fully connected neural network.

Fusion Approaches

Apart from different approaches to feature extraction from CT scans, different approaches to the fusion of different modalities were also possible. For each multimodal model, intermediate and late fusion were tested. Early fusion was not employed as concatenating the raw data was not meaningful for these modalities; flattening the CT data would result in an excessively large input feature vector and lead to the loss of spatial information.

The intermediate fusion approach for the 'Inception-v3 Transfer' model could also be considered early fusion, as the features extracted from the pre-operative CT scan were concatenated with the raw impedance matrix and cochlear shape data. However, due to the use of gradual unfreezing of the pre-trained model during training, we determined that in this case it aligned more closely with the definition of intermediate fusion.

3.3. Model Training

The models were trained using a supervised learning approach, where the LID of the most basal electrode was the target variable. The performance of each model was evaluated using 10-fold cross-validation to get a robust view of performance. This meant that for two of the folds, the models were trained on 127 cases and validated on 15 cases; and for the rest of the folds they were trained on 128 cases and validated on 14 cases.

For the experiments to be reproducible, all seeds were set manually. This ensured that each fold of the validation process had identical training and validation datasets over different runs and different experiments. It also ensured that weights were updated the same way each time for the same training run.

The training process involved minimizing a Mean Squared Error loss function using the Adam optimizer [75] with a dynamic learning rate scheduler. For multimodal transfer learning approaches, gradual unfreezing was used: the pre-trained model initially started with all weights frozen but was gradually unfrozen layer-by-layer as the training continued. Early stopping was used to stop training if validation loss had not decreased for a specified number of epochs, preventing overfitting and the waste of computational resources. After each training run, performance metrics for that fold were calculated using the weights of the model from the epoch with the smallest validation loss. The performance of the model was then estimated using the medians and standard deviations of the metrics over all 10 folds of the cross-validation.

The performance metrics were Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. The multimodal model with the lowest median RMSE was selected as the best model. MAE is the average absolute difference between predicted and true values, treating all errors equally; whereas RMSE squares the errors and thus penalizes larger differences more significantly compared to MAE, making it more sensitive to outliers. R^2 score, in contrast, explains how well the model explains the variance in the data, with values closer to 1 indicating better predictive power.

For classification performance analysis, a model with the same architecture as the best model was trained again using a 90%-10% training-validation split on the dataset, with 127 cases in the training dataset and 15 cases in the validation dataset. The model was trained on the regression task and the class predictions were made based on the predicted LID.

Hyperparameters were chosen through iterative experimentation. The performance of different configurations was evaluated using RMSE. Various combinations of hyperparameters were tested for each model and adjustments were made based on empirical observations to maximize performance.

3.4. Ablation Studies

The term ablation refers to the process of systematically altering or removing certain model components or inputs to evaluate their influence on performance. After the best multimodal model was selected, we ran ablation studies using that model to analyze how each of the input modalities influenced the performance. We used three different ablation methodologies to make the result analysis more robust.

1. **Exclusion ablation** where the modality was completely removed as an input from the model. This approach necessitated a slight change of architecture for the model for each modality, but the change was not severe as the feature extractor for that specific modality was removed, and the number of input nodes for the final fully connected neural network was changed.

2. **Random ablation** where the selected modality was replaced with Gaussian noise (mean $= 0$ and standard deviation $= 1$) with the same shape as the modality. The actual means and standard deviations of each modality based on our dataset were not used on purpose to prevent any kind of information leak from the selected modality.

3. **Zero-out ablation** where the selected modality was replaced with zeros in the same shape as the input modality. This effectively removed all information the model could extract from the selected modality.

```
if self.ablate == "impedance":
    if self.ablation_method == "random":
        imp_mat = torch.rand((12, 12))
    elif self.ablation_method == "zero":
        imp_mat = torch.zeros((12, 12))
else:
    imp_mat, ear_side = get_imp_mat(patient_dir)
    imp_mat = torch.Tensor(imp_mat)
```

Listing 3.1: Implementations of random and zero-out ablation.

Listing 3.1 shows the implementation of random and zero-out ablation for impedance matrices, the other modalities have identical implementations.

3.5. Experimental Setup

All models were implemented and all experiments were run using Python v3.12.7 and PyTorch v2.5.1 for CUDA v12.4. The CT scan preprocessing pipeline used functions from the MONAI v1.4.0 package. All experiments were conducted on a workstation equipped with two NVIDIA RTX™ A5000 GPUs, an AMD Ryzen™ Threadripper™ PRO 5955WX 16-core CPU, and 256 GB of RAM.

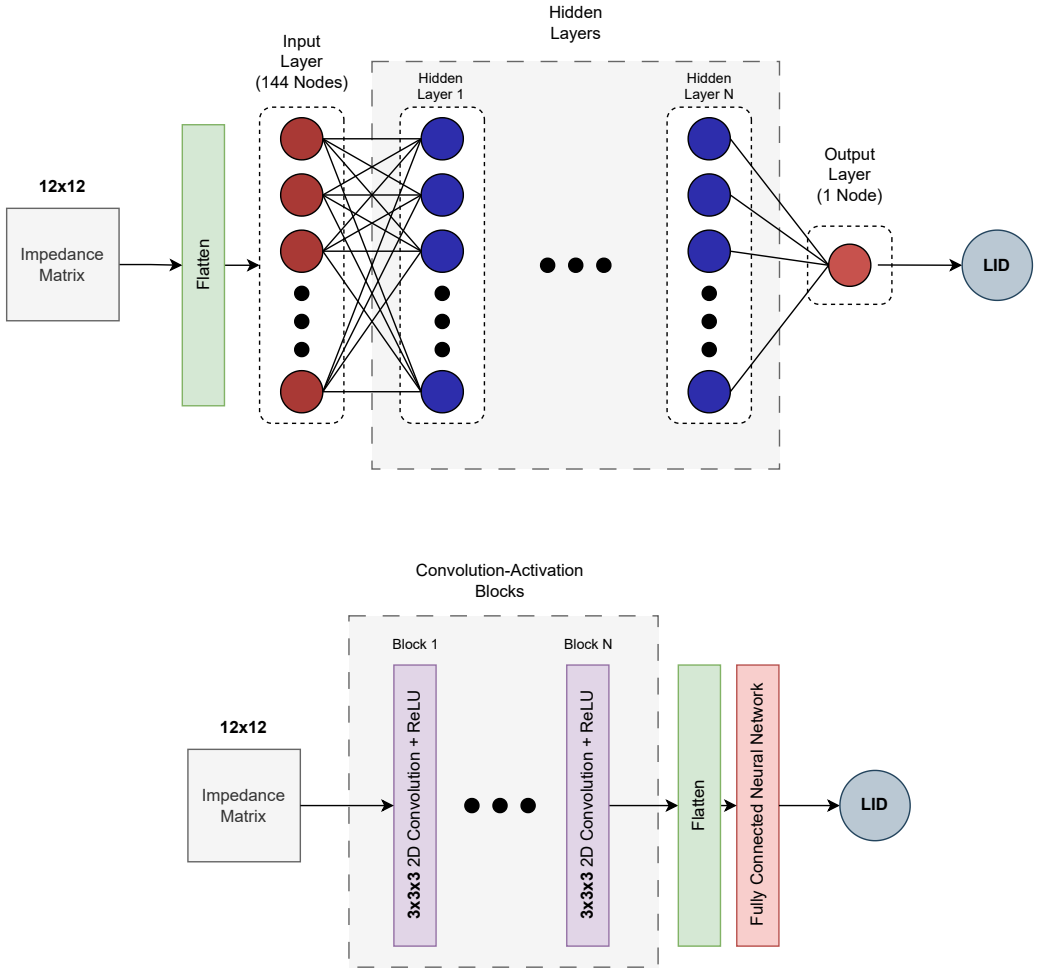


Figure 3.3.: Diagrams for the two general architectures for models with only impedance matrices as input. The top diagram shows the architecture for the single- and multi-layer perceptron approach, where the number of hidden layers and the number of nodes in these layers were configurable. The bottom diagram shows the architecture for the CNN based approach, where the number of convolution-activation blocks and the number of filters for each of the convolution layers was configurable.

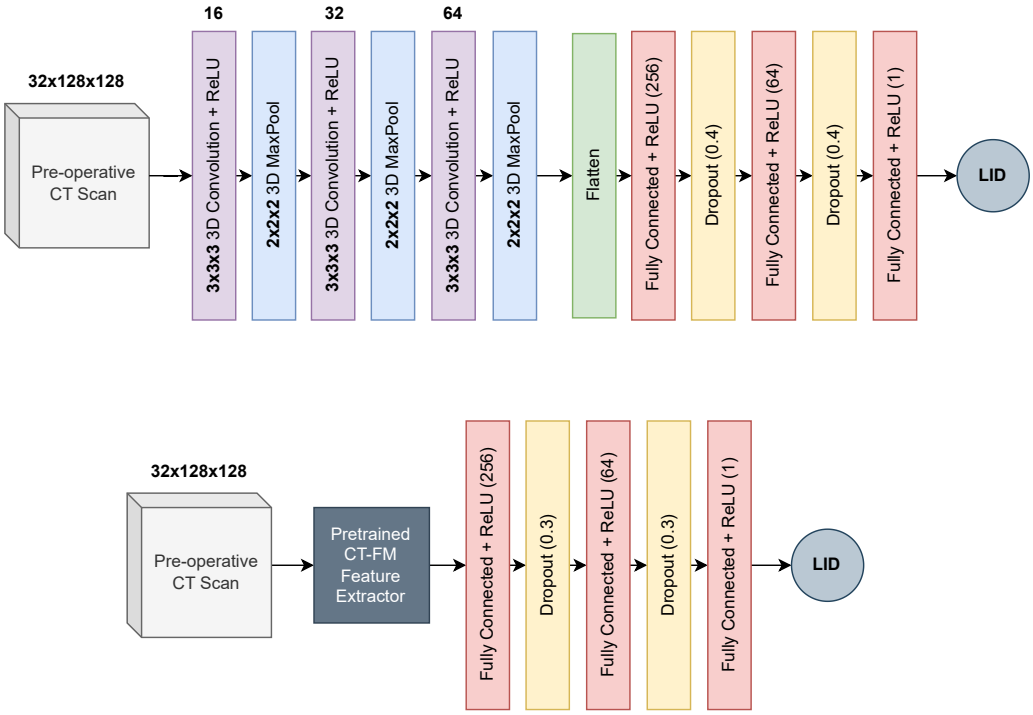


Figure 3.4.: Diagrams for the two architectures for models with only pre-operative CT scans as input. The top diagram shows the model using a 3D-CNN that is trained from scratch. The bottom diagram shows the model that uses a pre-trained CT-FM feature extractor. The numbers on top of the convolution-activation layers represent the number of filters. The numbers in fully connected layers represent the number of output features, and the number in the dropout layers represent the fraction of neurons that are randomly deactivated.

4. Results and Analysis

4.1. Dataset Analysis

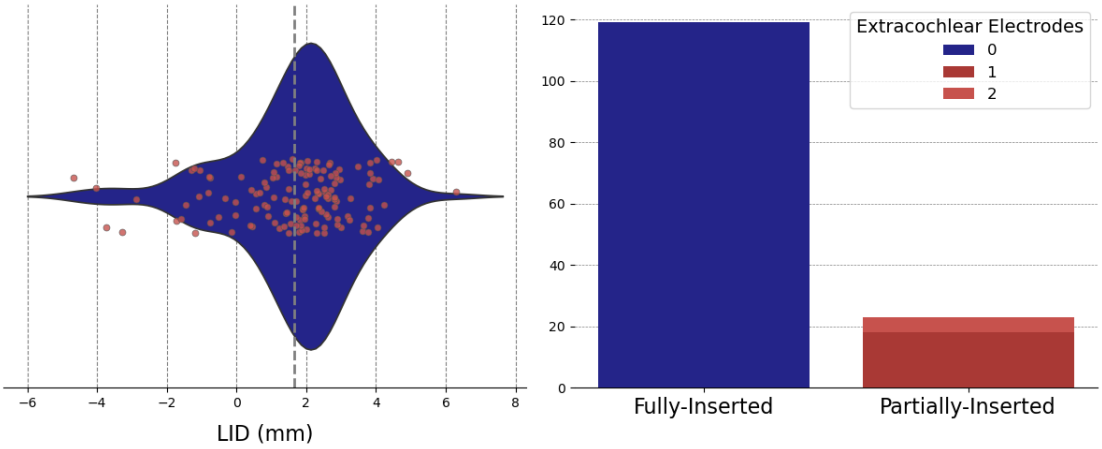


Figure 4.1.: LID distribution (left) and insertion classification distribution (right) of the dataset, consisting of 142 cases, that was used in this study. 119 fully-inserted and 23 partially-inserted cases are present. The gray line on the left plot shows the mean LID. LID: Linear Insertion Depth

This section provides an overview of the dataset used in this study. The dataset included 120 unique patients, of whom received a CI in both ears.

The mean LID was 1.65 mm and the standard deviation was ± 1.8 mm. The minimum and maximum LIDs were -4.68 mm and 6.31 mm respectively. Only 23 out of 142 (roughly 16%) of cases were partially-inserted, meaning the data was heavily skewed towards positive LIDs and fully-inserted cases. Of the 23 partially-inserted cases, 18 had a single extracochlear electrode and the remaining 5 had two extracochlear electrodes. Figure 4.1 shows the distribution of the ground truths from the perspective of regression and classification tasks.

The mean cochlear base length (a value) and cochlear base width (b value) were 9.15 mm and 6.98 mm, and the standard deviations were 0.47 mm and 0.46 mm. The minimum and maximum values for cochlear base length were 8.00 mm and 10.27 mm;

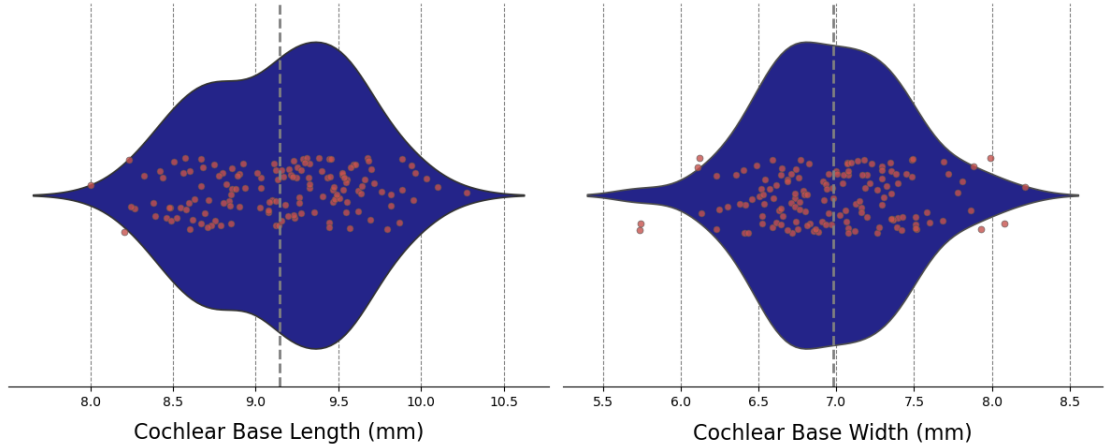


Figure 4.2.: Distribution of cochlear base length and width parameters of the dataset, consisting of 142 cases, that was used in this study. The gray dashed lines represent the means.

and for width, they were 5.74 mm and 8.21 mm. Figure 4.2 shows the distribution of these values.

4.2. Single Modality Model Results

In this section, we present the results of single-modality models that were trained on only one of the three modalities available in the dataset. The primary performance metric was RMSE. The model performances were analyzed based on a 10-fold cross-validation scheme. Lower values for RMSE and MAE indicate better performance, whereas a R^2 score closer to 1.0 suggests that the model explains more variance in the target variable. A Mean Model (MM) that predicted LID based on the mean LID of the training dataset was used as a baseline for comparison. As the MM does not explain any of the variance, it always had a R^2 score of 0.

4.2.1. Cochlear Shape Models

Each model was trained only on the two cochlear shape parameters, base length (a value) and base width (b value). Five models were trained in total, each with a different number of hidden layers and nodes. Table 4.1 shows the configurations and performance metrics for each of these models.

The best-performing model was found to be the 1-layer model with only a single

Table 4.1.: Performance comparison of models taking only cochlear shape parameters as input. The number of hidden layers and the number of nodes in each layer are given for each different Multilayer Perceptron (MLP) configuration. Model performances are compared against the baseline mean model. The best performance is shown in bold. MLP: Multilayer Perceptron; MM: Mean Model.

Model	**Layers**	**Nodes**	**Median Performance Metrics**		
			RMSE (mm)	MAE (mm)	R^2 Score
Single Layer	0	-	1.66	1.19	0.01
MLP	1	16	**1.62**	1.20	0.03
	2	16, 8	1.62	1.21	0.02
	3	16, 8, 4	1.64	1.21	0.00
	4	16, 8, 4, 2	1.66	1.25	0.00
MM (Baseline)	-	-	1.73	1.31	0.00

hidden layer. However, this model performed only marginally better than the 2-layer model with two hidden layers and the baseline MM. The median RMSE was reduced by 0.4% and 6.4% over these two models, respectively. The single layer model performed almost identically to the MM, while the deeper 3-layer and 4-layer models showed decreasing performance.

The 3-layer and 4-layer models both had a median R^2 score of 0, identical to the MM. One interesting observation was that the median MAE of the single layer model was lower than the best-performing model, while its median RMSE was larger.

4.2.2. Impedance Matrix Models

Table 4.2.: Performance comparison of perceptron models taking only impedance matrices as input. The number of hidden layers and the number of nodes in each layer are given for each different MLP configuration. Model performances are compared against the baseline mean model. The best performance is shown in bold. MLP: Multilayer Perceptron; MM: Mean Model.

Model Type	**Layers**	**Nodes**	**Median Performance Metrics**		
			RMSE (mm)	MAE (mm)	R^2 Score
Single Layer	0	-	1.29	1.06	0.40
MLP	1	64	1.24	1.03	0.44
	2	64, 32	1.20	1.03	0.50
	3	64, 32, 16	**1.13**	0.96	0.43
	4	64, 32, 16, 8	1.18	1.01	0.44
MM (Baseline)	-	-	1.73	1.31	0.00

Table 4.3.: Performance comparison of convolution-based models taking only impedance matrices as input. The number of convolution layers and the number of filters in each layer are given for each different CNN configuration. Model performances are compared against the baseline mean model. The best performance is shown in bold. CNN: Convolutional Neural Network; MM: Mean Model.

Model Type	**Layers**	**Filters**	**Median Performance Metrics**		
			RMSE (mm)	MAE (mm)	R^2 Score
CNN	1	8	1.49	1.16	0.04
	2	8, 16	1.38	1.03	0.21
	3	8, 16, 32	**1.23**	1.04	0.31
	4	8, 16, 32, 64	1.42	1.12	0.32
MM (Baseline)	-	-	1.73	1.31	0.00

Each model was trained using only the impedance matrices as input. As two approaches were possible for this, two different categories of models were tested. Table 4.2 shows the results for the models based on the first approach, where the impedance

matrices were flattened and fed into perceptron models as feature vectors. Table 4.3 shows the results of models from the second approach, where the impedance matrices were treated as images and fed through CNNs.

The performances of all models in the first approach were higher than the MM. Even the worst-performing model, a single layer perceptron, achieved a 25.4% reduction in median RMSE compared to the MM. Median RMSE decreased as the depth of the models increased; with the 1-, 2-, and 3-layer MLP models achieving 28.3%, 30.6%, and 34.7% reductions in median RMSE compared to the MM, respectively. However, increasing depth beyond three layers had a negative impact on performance, as the 4-layer MLP achieved a 31.8% decrease in median RMSE over the MM, compared to the 3-layer MLP model's 34.7% decrease. For this approach, the MAE followed the same performance ranking as RMSE. Even though the 3-layer MLP model was the best-performing model based on RMSE, it had a lower median R^2 score than every other model except the MM and the single layer perceptron model.

The second approach also produced models that outperformed the MM. All models created using this approach outperformed the MM in every metric. As with the first approach, the performance gains got larger as the models got deeper but degraded after a certain depth. The 1-, 2-, and 3-layer CNNs had 13.9%, 20.2%, and 28.9% lower median RMSE than the MM, respectively. The 4-layer CNN broke this pattern by having a median RMSE 15.4% higher than the 3-layer CNN, and a 7.7% higher median MAE. However, the R^2 score continued to increase as the model depth increased.

Looking at both approaches together, the best model from the first approach (3-layer MLP) achieved a 8.1% lower median RMSE compared to the best model from the second approach (3-layer CNN). It also had a 7.7% lower median MAE and a median R^2 score that was 0.12 higher.

4.2.3. CT Models

Table 4.4.: Performance comparison of models using different feature extractors, taking only pre-operative CT scans as input. Model performances are compared against the baseline mean model. The best performance is shown in bold. CNN: Convolutional Neural Network; CT-FM: Computed Tomography - Foundation Model; MM: Mean Model.

Feature Extractor	**Median Performance Metrics**		
	RMSE (mm)	MAE (mm)	R^2 Score
3D-CNN	**1.61**	1.21	0.01
Pre-trained CT-FM	1.61	1.23	0.08
MM (Baseline)	1.73	1.31	0.00

The two models in this category were trained only on the pre-operative CT scans. Table 4.4 shows the results of both models.

Results showed marginally better performances than the MM. Both models had a median RMSE only 6.9% lower than the MM. The model with the 3D-CNN feature extractor showed marginally better performance in terms of median MAE, achieving 1.6% and 7.6% lower median MAE compared to the model with the CT-FM feature extractor and the MM, respectively. However, the model with the CT-FM feature extractor showed much better relative performance based on R^2 scores, with a score of 0.08 compared to the other model's 0.01.

4.3. Multimodal Models

In this section, we present the results and selection of the best multimodal model that took all three modalities (pre-operative CT scans, impedance matrices, and cochlear shape parameters) as input. The MM was used as a baseline.

Table 4.5.: Performance comparison of multimodal models taking all three modalities as input. For each model, intermediate and late fusion configurations are shown. Out of the five models, the first three were trained fully from scratch (denoted by *), while the last two ('Transfer' models) used a pre-trained model as a feature extractor for the CT scans. Model performances are compared against the baseline mean model. The best performance is shown in bold. CNN: Convolutional Neural Network; CT-FM: Computed Tomography - Foundation Model; MM: Mean Model.

Model	**Fusion Method**	**Median Performance Metrics**		
		RMSE (mm)	MAE (mm)	R^2 Score
Shallow CNN*	Intermediate	**1.14**	0.95	0.47
	Late	1.35	1.04	0.22
Deep CNN*	Intermediate	1.27	1.05	0.33
	Late	1.64	1.25	0.01
GoogLeNet3D*	Intermediate	1.49	1.11	0.07
	Late	1.43	1.08	0.02
Inception-v3 Transfer	Intermediate	1.25	0.97	0.40
	Late	1.31	0.99	0.12
CT-FM Transfer	Intermediate	1.35	1.01	0.26
	Late	1.51	1.12	0.36
MM (Baseline)	-	1.73	1.31	0.00

All multimodal models were configured to have either intermediate or late fusion.This was done without significantly altering the model architectures to avoid introducing new variables for the performance. Table 4.5 and Figure 4.3 show the results for these models. Early fusion was not considered as flattening the raw CT scan data would result in excessively large input vectors and would lead to the loss of spatial information.

The performance metrics across all models showed considerable variation. The

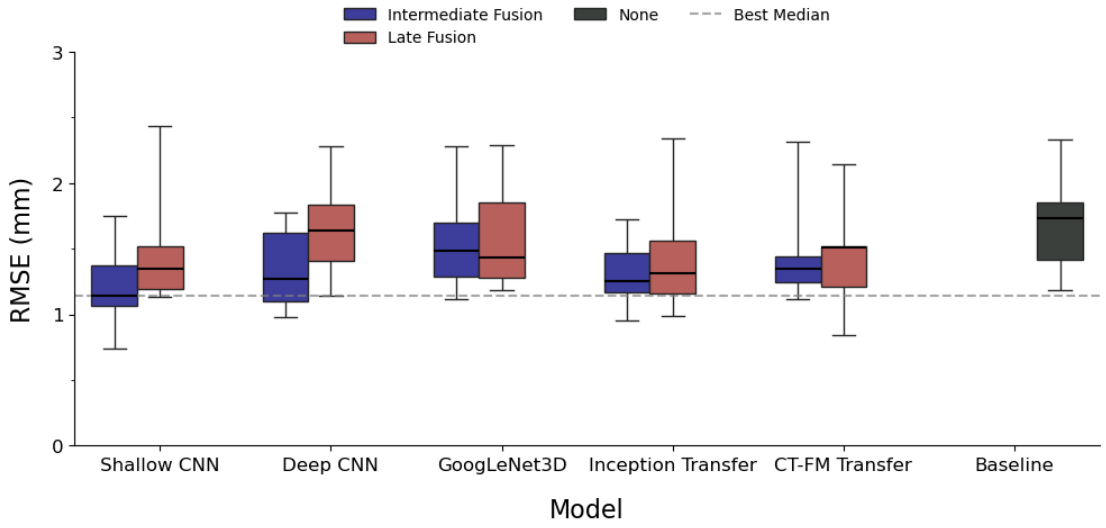


Figure 4.3.: Comparison of multimodal models, each with two different fusion approaches, based on RMSE distribution across a 10-fold cross-validation scheme. The dashed line represents the best median RMSE across all models. The different fusion approaches are color-coded. The baseline is the mean model. CNN: Convolutional Neural Network; CT-FM: Computed Tomography - Foundation Model.

median RMSE ranged from 1.14 mm (best) to 1.64 mm (worst), the median MAE ranged from 0.95 mm (best) to 1.25 mm (worst), and the median R^2 score ranged from 0.47 (best) to 0.01 (worst).

The best-performing model across all metrics was the Shallow CNN model with intermediate fusion. This model showed a 34.1% lower median RMSE, 27.5% lower median MAE, and an increase of 0.47 in median R^2 score compared to the MM. Compared to the second-best-performing model based on RMSE, the Inception-v3 Transfer model with intermediate fusion, the Shallow CNN model with intermediate fusion had an 8.8% lower median RMSE.

The worst-performing model based on median RMSE was the Deep CNN model with late fusion. It had a median RMSE that was only 5.2% lower than the MM. However, the Deep CNN model with intermediate fusion was the third best-performing model based on the same metric, with a median RMSE 26.6% lower than the MM.

The intermediate fusion approach yielded a lower median RMSE on almost every model, except for the GoogLeNet3D model which had a 4% lower median RMSE using the late fusion approach compared to the intermediate fusion approach. Across all

models, the median RMSE for intermediate fusion ranged from 1.14 mm to 1.49 mm, while for late fusion it ranged from 1.31 mm to 1.64 mm. The average decrease in median RMSE was 13.3% for the models where intermediate fusion had a lower median RMSE compared to late fusion. The biggest decrease in median RMSE from late to intermediate fusion was for the Deep CNN model, which exhibited a 22.6% decrease.

Figure 4.3 shows that models with intermediate fusion exhibited lower variance in RMSE compared to models with late fusion. Shallow CNN with intermediate fusion exhibited relatively low variance, further cementing its place as the best model. Both transfer models as well as GoogLeNet3D showed significant overlap in RMSE distribution between their intermediate and late fusion varieties.

4.4. Modality Ablation

We conducted an ablation study on the best-performing multimodal model, Shallow CNN with intermediate fusion, to analyze the contributions of each modality to model performance. We used three different ablation methods for each modality, which are explained in detail in Section 3.4.

Table 4.6.: Performance comparison of three different ablation methods applied to the best performing multimodal model (Shallow CNN with intermediate fusion). For each ablation method, the results for applying it to each of the three modalities are shown. The results without any ablation are given as a baseline. The performance closest to the baseline is shown in bold. CT: Computed Tomography.

Ablation Method	**Ablated Modality**	**Median Performance Metrics**		
		RMSE (mm)	MAE (mm)	R^2 Score
Exclusion	CT Scans	1.34	0.99	0.30
	Impedance Matrices	1.76	1.31	-0.02
	Shape Parameters	1.27	0.98	0.29
Random	CT Scans	1.28	0.98	0.32
	Impedance Matrices	1.75	1.30	-0.01
	Shape Parameters	1.31	0.99	0.33
Zero-out	CT Scans	1.29	0.98	0.35
	Impedance Matrices	1.73	1.31	0.01
	Shape Parameters	**1.18**	0.98	0.49
None (Baseline)	-	1.14	0.95	0.47

Table 4.6 and Figure 4.4 show the results of this study. Ablating the impedance matrices consistently had the most impact on performance metrics. The median RMSE across all ablation methods for impedance matrices was 1.75 mm, signifying a 53.5% increase over the baseline. Meanwhile, the median MAE was 1.31 mm, a smaller yet still substantial 37.9% increase. The R^2 score was 0.48 smaller on average, a decrease of more than 100%. R^2 score was consistently near zero for all methods, signifying little to no predictive power.

The second most impactful modality was CT scans, having higher median RMSE compared to the cochlear shape parameters for two of the ablation methods. However, for the exclusion method, cochlear shape parameters had more of an effect on perfor-

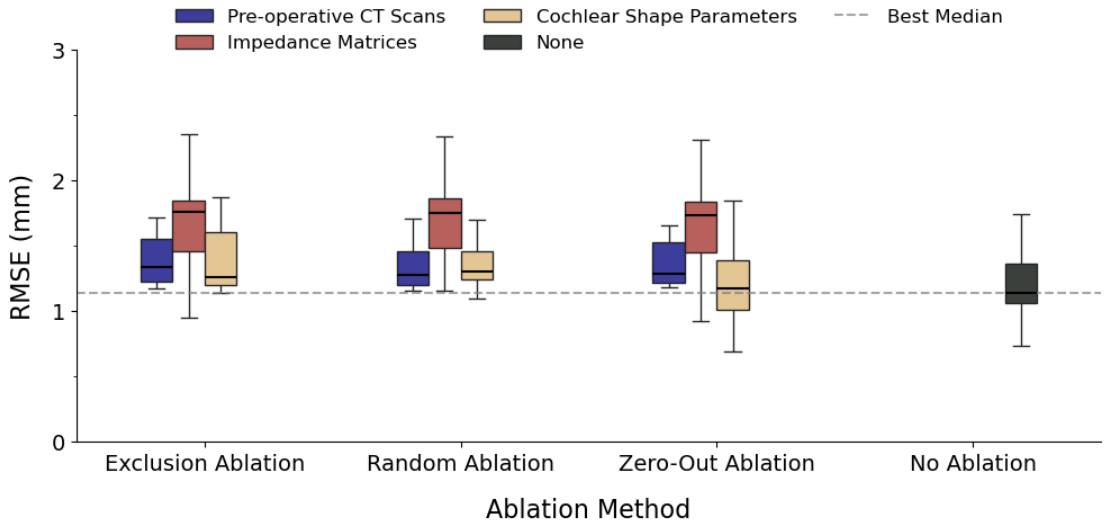


Figure 4.4.: Comparison of ablation results, based on RMSE distribution across a 10-fold cross-validation scheme. Results are grouped by ablation method. The ablated modalities are color-coded. The model with no ablation is included as a baseline. The dashed line represents the best median RMSE.

mance. The model still retained diminished yet substantial performance without the CT scans. The median RMSE across all methods was 1.29 mm, a 13.2% increase over the baseline. The MAE was 0.98 mm, a small 3.2% increase over the baseline; while the median R^2 score was 0.32, a 0.15 decrease and only 68.1% of the baseline.

The modality with results closest to the baseline were the cochlear shape parameters. The average median RMSE, MAE, and R^2 score was 1.27 mm, 0.98 mm, and 0.33, respectively. These signify an overall increase of 11.4% and 3.2% for RMSE and MAE, respectively; and a 0.14 decrease for R^2 score over the baseline. Using zero-out ablation produced a median R^2 score 0.02 higher than the baseline.

Ablating each modality had similar results for each of the ablation methods, showing a tight spread across all metrics. Overall, the exclusion method had the most drastic performance drops compared to the other methods.

4.5. Classification

A model with the best-performing multimodal architecture based on the regression task, Shallow CNN with intermediate fusion, was trained on a random 90% - 10% training-validation split of the dataset. Both datasets had similar proportions of partially-inserted cases: 15.6% (20 out of 127) in the training set and 20% (3 out of 15) in the validation set. The model was trained for the regression task, and class predictions were made from the predicted LID values based on whether the value was positive (fully-inserted) or negative (partially-inserted).

K-fold cross-validation was not used for the classification metrics due to the low number of extracochlear cases in the dataset (Figure 4.1). In many folds, the validation dataset contained few or no extracochlear cases, leading to unreliable generalization of the classification metrics.

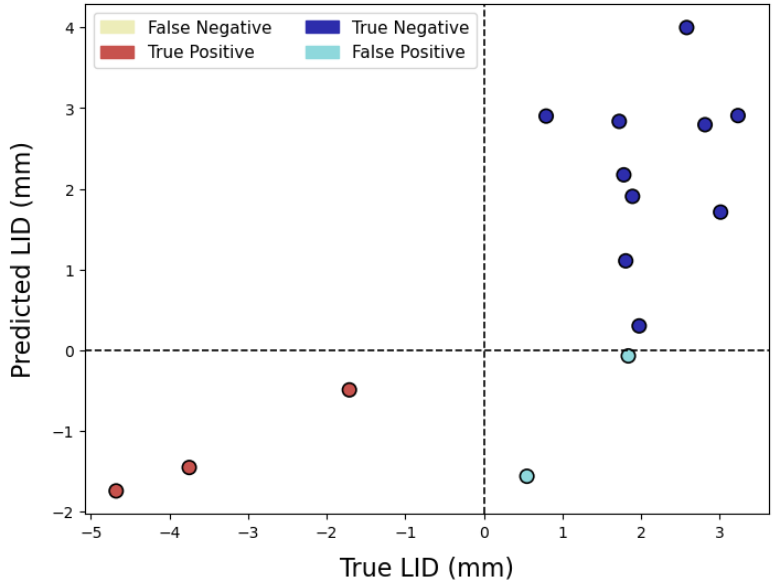


Figure 4.5.: Scatter plot showing predicted vs. true LID values, color-coded for classification outcomes for 15 cases. The positive class is partially-inserted, and the negative class is fully-inserted. The dashed lines indicate classification thresholds. The vertical line ($X = 0$) separates the true class labels (left: negative, right: positive), while the horizontal line ($Y = 0$) separates the predicted class labels (below: negative, above: positive).

The predicted vs. true LID values of the validation set for this training run and the corresponding classification outcomes are given in Figure 4.5. The correspond-

Table 4.7.: Confusion matrix showing true vs. predicted class labels. Positive class is partially-inserted, negative class is fully-inserted. TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative.

		True Class	
		Positive	Negative
Predicted Class	Positive	TP = 3	FP = 2
	Negative	FN = 0	TN = 10

Table 4.8.: Classification performance metrics based on the confusion matrix (Table 4.7). AUC: Area Under the Receiver-Operating Characteristic Curve.

Metric	**Value**
Accuracy	86.67%
Precision	60.00%
Recall	100.00%
F1-score	75.00%
AUC	91.67%

ing confusion matrix and performance metrics are given in Table 4.7 and Table 4.8, respectively.

The validation dataset contained 3 partially-inserted, or positive, cases. The model was able to classify all of these correctly and achieved a perfect recall score. No false negatives were present. Of the 12 fully-inserted, or negative, cases in the dataset, the model only mislabeled 2.

The model had high accuracy and AUC scores, suggesting it was efficient at differentiating between the two classes. The perfect recall score value demonstrated that the model successfully identified all positive cases.

5. Discussion

5.1. Single Modality and Multimodal Models

The single-modality models based only on the cochlear shape parameters exhibited performances very similar to the MM (Table 4.1). This suggested that the cochlear base length and width were not good predictors for LID values, especially based on the R^2 scores. Looking at their Spearman's rank correlation values with LID, both parameters showed weak to moderate but statistically significant positive correlations ($\rho = 0.29$, $p = 0.00039$ for length and $\rho = 0.21$, $p = 0.01$ for width), further corroborating this conclusion.

The models based only on the pre-operative CT scans did not perform any better, but the model that used the pre-trained CT-FM feature extractor showed a marginally better R^2 score (Table 4.4). Between these two modalities, CT scans were a better predictor for LID values.

Results from the impedance matrix-only models were drastically better (Tables 4.2 and 4.3), suggesting significant predictive power. This observation is in line with existing research [23]–[30].

The performance of multimodal models ranged drastically from substantial to almost indistinguishable from the baseline MM (Table 4.5). Overall, the intermediate fusion approach showed better results than late fusion, indicating that capturing the relationships between the modalities in a joint feature representation was more effective at predicting LID values than separate feature representations for each modality. This was the expected result, as the pre-operative CT scans and cochlear shape parameters by themselves showed poor predictive power (Tables 4.1 and 4.4).

The Shallow CNN model with intermediate fusion, trained completely from scratch, was the best-performing multimodal model for each of the result metrics. Compared to most other multimodal models, this model concatenated the raw impedance matrix and cochlear shape parameters with features extracted from the pre-operative CT scans. This might have enabled better joint representations by allowing the model to learn the relationship between impedance values and features extracted from the CT scan rather than relying on high-level fused features. The Shallow CNN model outperformed the Deep CNN model on all metrics. The Shallow model extracted a significantly higher number of features from the CT scan than the Deep model (see Figures A.1 and

A.2), suggesting that more complex representations of the CT scans may be beneficial for performance. The Shallow CNN also outperformed both transfer learning-based models, suggesting that the features extracted by the pre-trained models were not sufficiently informative for predicting LID.

The CT-FM Transfer model performed worse than the Inception-v3 Transfer model, which was not expected as the CT-FM was originally trained on CT scans of the body, including the head, whereas the Inception-v3 model was trained on color images of a wide range of objects (animals, everyday objects, vehicles, etc.). We interpreted this as an indication that the features extracted by the CT-FM were less transferable for the LID estimation task than initially expected; possibly as they focused more on high-level anatomical structures, and our task required finer, more localized features. In contrast, the Inception-v3 Transfer model performed the second best out of all multimodal models, which suggested that it may extract features that are more adaptable for this task. This unexpected result suggested that pre-training on CT scans did not guarantee superior performance for this task, however further fine-tuning on more domain-specific CT scans showing the cochlea could help improve the transferability.

The best-performing model based only on impedance matrices, the 3-layer MLP, performed marginally better than the multimodal Shallow CNN model based on median RMSE, 1.13 mm vs. 1.14 mm. However, the Shallow CNN had a better median MAE and R^2 score; 0.95 mm vs. 0.96 mm and 0.47 vs. 0.43, respectively. These metrics suggested that the models predicted LID with comparable accuracy, but the Shallow CNN provided slightly more consistent predictions. This might have been caused by the inclusion of the two additional modalities, CT scans and cochlear shape parameters, reducing the reliance on any one modality and leading to more stable predictions.

5.2. Ablation Studies

The ablation study we conducted demonstrated valuable results on the contribution of each modality on the Shallow CNN model's performance (Table 4.6). Consistent with the findings from the single modality model results, impedance matrices contributed the most to the model's predictive power, followed by pre-operative CT scans, and lastly cochlear shape parameters. None of the results showed equal or lower median RMSE than the original model with no ablation, suggesting that each modality contributed to the performance, justifying a multimodal approach.

5.3. Classification

The classification performance analysis was not as robust as the regression performance analysis as it was only based on a single training run (Table 4.8). The model achieved a perfect recall score, which was not realistic for generalized performance, but suggested that it was highly sensitive to detecting positive (partially-inserted) cases. This came at the cost of precision as the model misclassified some negative (fully-inserted) cases, leading to a precision score of 60.0%. Consequently, the F1-Score of 75.0% reflected an acceptable balance of precision and recall. The high AUC score of 91.67% suggested that the model had high discriminative ability, meaning it could effectively distinguish between fully- and partially-inserted cases overall. As the dataset was skewed towards fully-inserted cases (Figure 4.1), the F1-Score was a better representation of model performance compared to AUC. These results suggested that this model could aid in detecting partially-inserted CI cases in clinical practice. This would reduce the amount of radiation patients are exposed to because of post-operative CT scans, while still reliably detecting the cases where a scan is necessary. However, a more rigorous validation of classification performance is needed to confirm these findings.

5.4. Comparison with State-of-the-Art Models

Our best model (Shallow CNN with intermediate fusion) predicted the LID of the most basal electrode with an absolute error of 0.98 mm $\pm$ 0.23 mm (mean $\pm$ standard deviation) over 10-fold cross-validation. The previous best model, based on Extremely Randomized Trees, achieved an absolute error of 0.8 mm $\pm$ 0.6 mm through leave-one-out cross-validation [23]. While our model exhibited a higher MAE, the lower standard deviation suggested that our predictions were more consistent across different test folds. However, the discrepancy in cross-validation strategies should also be considered when comparing these results. The classification results were not suitable to be compared as ours were based on a single run of validation, giving a poor view of general classification performance. Overall, the choice of the better model was not straightforward and depended on what aspect of performance was more important. The best model of [23] had better absolute accuracy, while ours performed more consistently and was, therefore, more predictable.

5.5. Limitations

We faced several limitations while conducting this study. Although the dataset comprised 142 cases—larger than those used in previous studies on the subject—it remained

limited, especially for a DL approach. As our main focus was creating multimodal models, the increase in the number of cases by an order of magnitude was intractable and therefore outside the scope of this study. There was also a significant class imbalance with 119 negative cases (fully-inserted) and 23 positive cases (partially-inserted), resulting in a 5.2 : 1 ratio. This imbalance made classification a less feasible primary focus. The pre-operative CT scans varied in resolution and this affected the annotation process, which might have led to less accurate ground truths for cases with lower-resolution scans and thus impacted model performance. Hyperparameter tuning was done empirically and could be improved with more formal methods (such as Bayesian optimization) to further improve the performance of the models.

5.6. Future Work

While we were able to create a multimodal model with performance comparable to the state-of-the-art, we were not able to set a new benchmark. Our results showed promising results for a multimodal approach for both the regression of the LID of the most basal electrode and fully- vs. partially-inserted CI classification. As such, future work should focus on further investigating multimodal models, especially intermediate fusion-based ones as they performed the best.

A broader range of model architectures, such as 3D vision transformers which have demonstrated promising results in 3D medical segmentation tasks [76], could be explored for improved CT feature extraction. Alternatively, the proposed pre-trained feature extractors, Inception-v3 and CT-FM could be further fine-tuned, or entirely different pre-trained feature extractors could be investigated. Additionally, a new feature extractor could be trained using unsupervised learning on a much bigger, external dataset of head CT scans.

Manual feature extraction from impedance matrices, as proposed by [23], could be investigated to gauge whether they translate to a multimodal approach. Further hyperparameter tuning, especially with regards to CT scan voxel dimensions, could be conducted to improve performance. The dataset could be expanded by adding new cases, generating artificial cases, or applying data augmentation.

6. Summary

This study explored the effectiveness of a multimodal approach employing pre-operative CT scans, impedance matrices, and cochlear shape parameters as input for predicting the LID of the most basal electrode in the CI electrode array. Various model architectures for both unimodal and multimodal models were evaluated. Intermediate and late fusion approaches for all multimodal models were tested. Transfer learning was employed for CT feature extraction and compared with training from scratch. The model performances were compared using 10-fold cross-validation and the best model was selected based on median RMSE.

The Shallow CNN model with intermediate fusion emerged as the best-performing multimodal model. While its median RMSE was slightly higher than the best-performing impedance matrix based unimodal model (0.01 mm), it outperformed the impedance-only model on median MAE and median R^2 score. To further confirm that each modality contributed to model performance, ablation studies using three different methods were conducted. The impedance matrices proved to be the most informative modality for the LID regression task, followed by the CT scans, and the cochlear shape parameters. No ablated model performed the same as or better than the non-ablated model, suggesting that all modalities contributed to the performance. The classification (fully- vs partially-inserted) performance of the model was measured using a single training run. This resulted in a perfect recall score (100%), but relatively lower precision (60%). The F1-Score (75%) highlighted the trade-off between recall and precision while still showing good predictive capability. However, further testing is necessary to more robustly evaluate classification performance.

The best model (Shallow CNN with intermediate fusion) performed comparably to the current best model from [23] that was based only on impedance matrices. The best model of this study had an absolute error of 0.98 mm $\pm$0.23 mm, while the best model of [23] had 0.8 mm $\pm$0.6 mm (mean $\pm$ standard deviation). This suggested that while the best model from this study had larger errors, it performed more consistently. The difference in validation strategies, 10-fold vs. leave-one-out cross-validation, should be considered when comparing these results.

While our results showed promise, further improvements could be made by refining model hyperparameters, increasing the size of the dataset, and using different model architectures.

7. Conclusion

We assessed the feasibility of a multimodal approach using pre-operative CT scans, impedance matrices, and cochlear shape parameters as a radiation-free approach for electrode localization. We used a dataset consisting of 142 cases, 23 of them being partially-inserted. Our best model, based on 3D-convolution layers for CT feature extraction and employing an intermediate fusion approach, achieved a submillimeter MAE and small variation across 10-fold cross-validation. It performed substantially better than the MM which served as the baseline, and the performance was comparable to state-of-the-art models. The classification metrics implied good performance but were based on a single training run and thus were not robust enough to draw general conclusions. An ablation study using three different methods for ablation showed that each modality contributed to the model performance, albeit at different magnitudes, justifying a multimodal approach.

A. Multimodal Model Diagrams

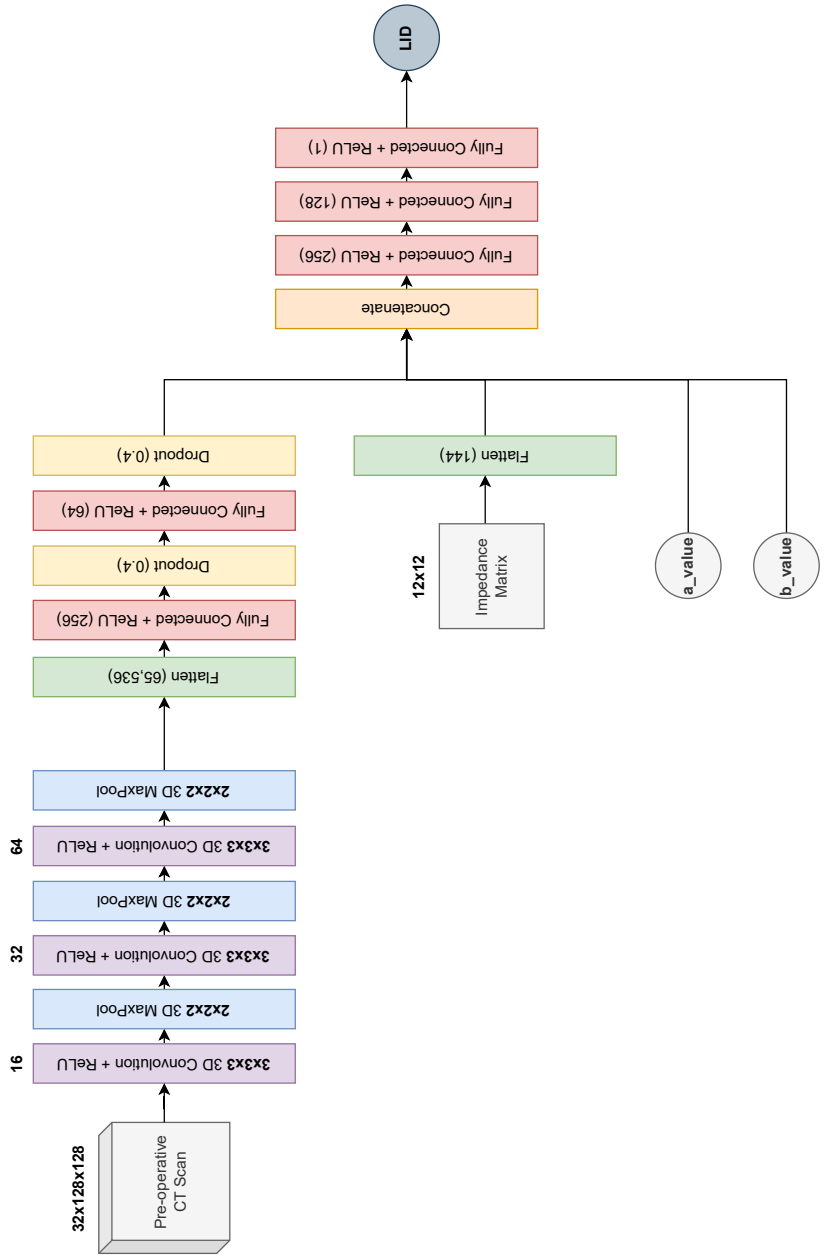


Figure A.1.: Diagram for the Shallow CNN model with intermediate fusion. The numbers on top of the convolution-activation layers represent the number of filters. The numbers in fully connected layers represent the number of output features, the number in the dropout layers represents the fraction of neurons that are randomly deactivated, and the number in the flatten operations represents the length of the output vector.

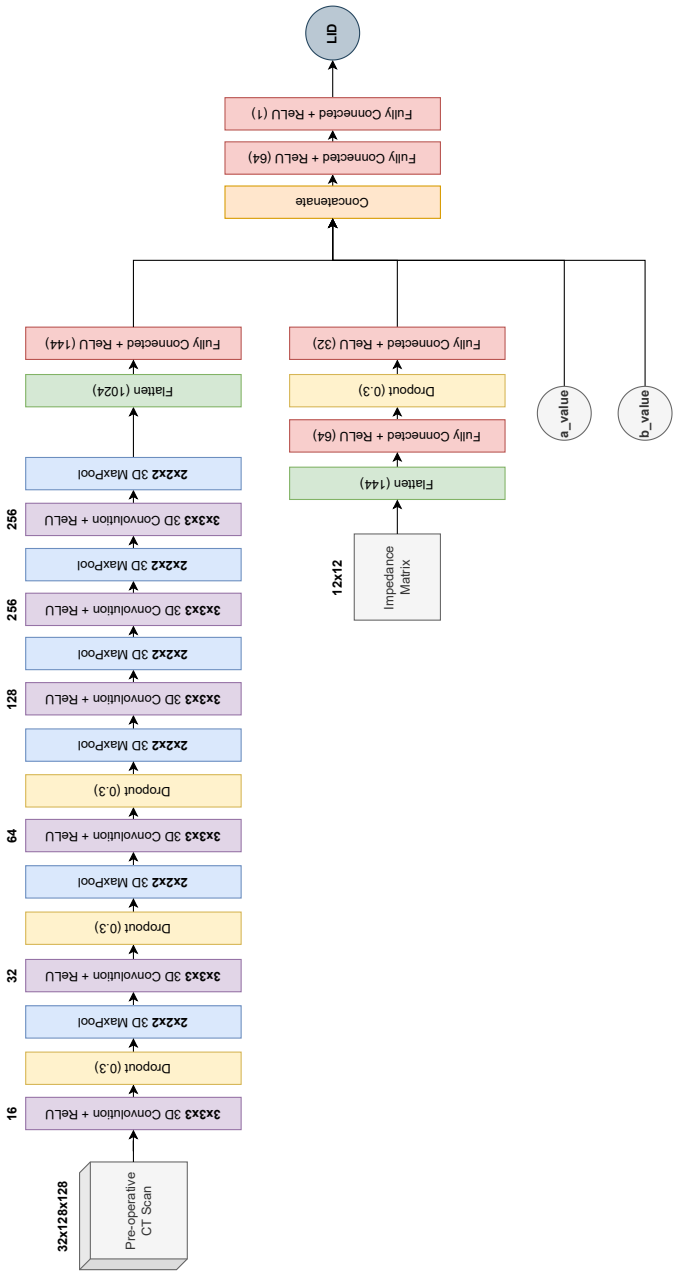


Figure A.2.: Diagram for the Deep CNN model with intermediate fusion. The numbers on top of the convolution-activation layers represent the number of filters. The numbers in fully connected layers represent the number of output features, the number in the dropout layers represents the fraction of neurons that are randomly deactivated, and the number in the flatten operations represents the length of the output vector.

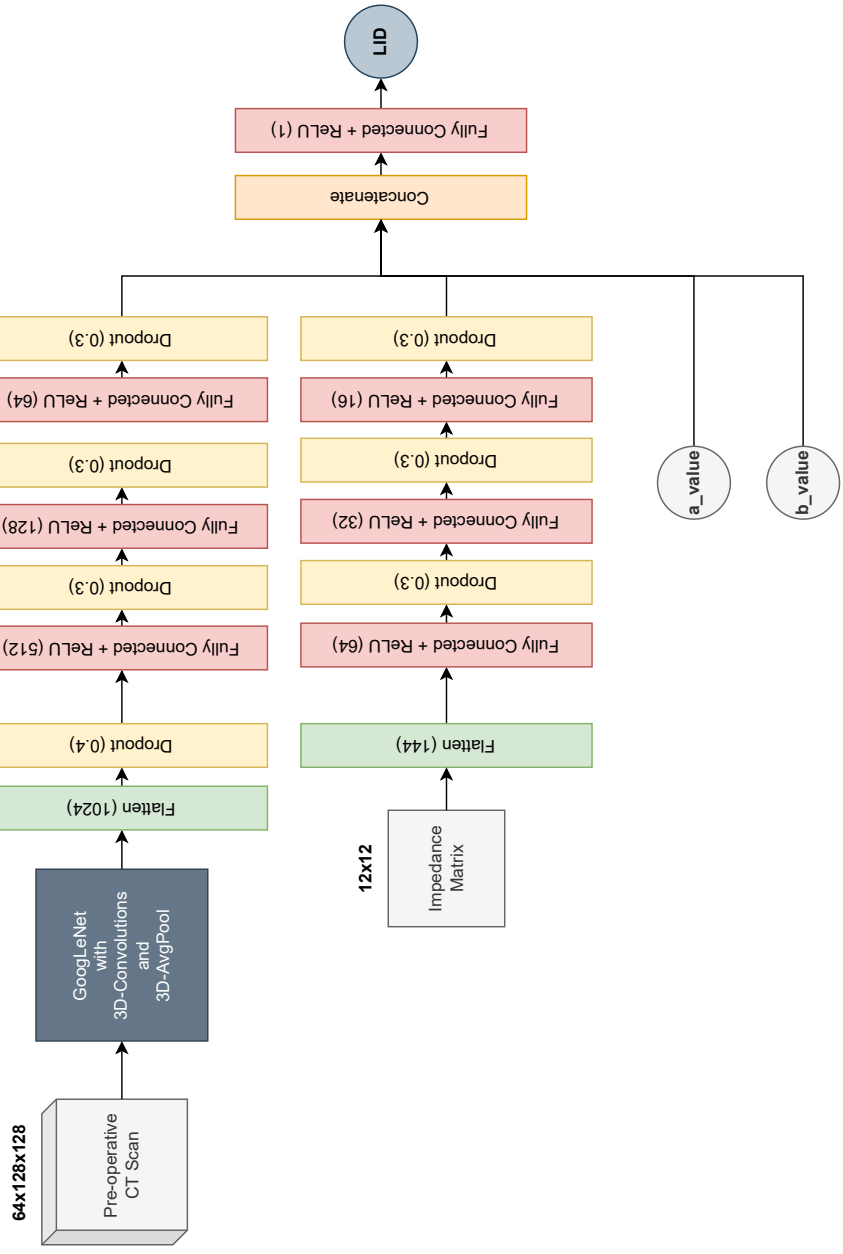


Figure A.3.: Diagram for the GoogLeNet3D model with intermediate fusion. The numbers in fully connected layers represent the number of output features, the number in the dropout layers represents the fraction of neurons that are randomly deactivated, and the number in the flatten operations represents the length of the output vector.

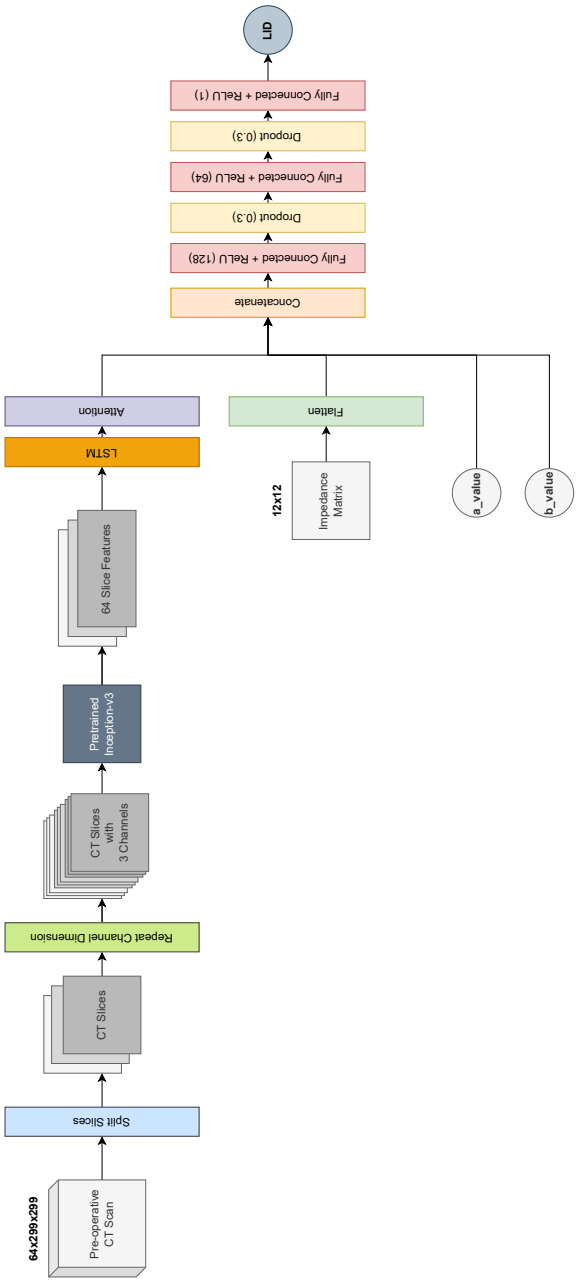


Figure A.4.: Diagram for the Inception-v3 Transfer model with intermediate fusion. The numbers in fully connected layers represent the number of output features, and the number in the dropout layers represents the fraction of neurons that are randomly deactivated. The LSTM module has a hidden size of 512 and a single hidden layer. LSTM: Long Short-Term Memory

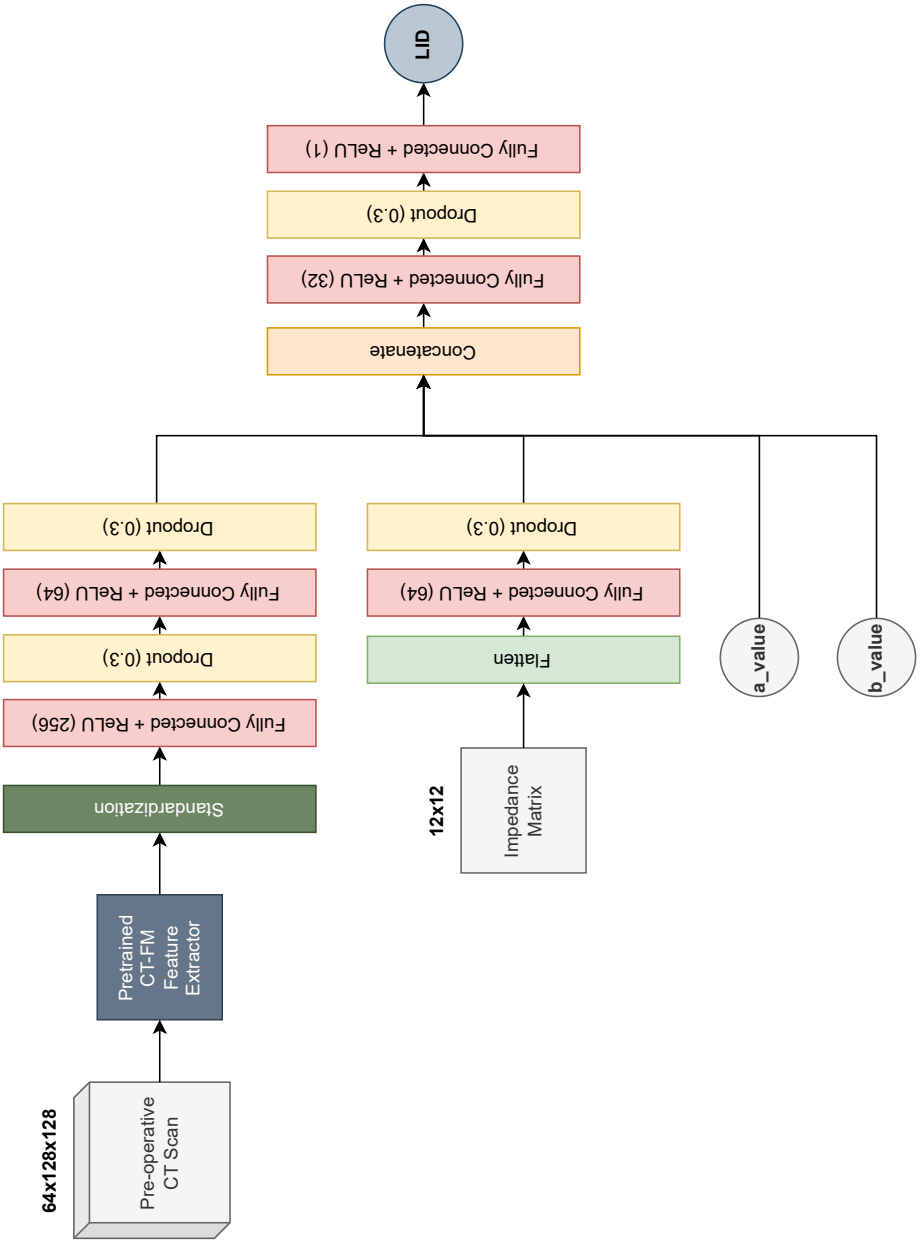


Figure A.5.: Diagram for the CT-FM Transfer model with intermediate fusion. The numbers in fully connected layers represent the number of output features, and the number in the dropout layers represents the fraction of neurons that are randomly deactivated. The standardization is done using pre-calculated means and standard deviations.

B. Code Snippets

```python
import json

import numpy as np
import pydicom as dicom

def crop_around_landmark(
    ds: dicom.dataset.FileDataset,
    pixel_arrays: np.array,
    patient_dir: str,
    landmark_name: str = "C",
    crop_size_cm: tuple[float, float, float] = (5.0, 5.0, 2.5),
    voxel_spacing: Optional[tuple[float, float, float]] = None,
    offset_cm: Optional[tuple[float, float, float]] = None,
    ear_side: Optional[str] = None,
) -> np.array:
    """
    Crop 3D Numpy array of CT data pixel arrays around given landmark
    with given crop size and offset.

    Args:
    - ds: First DICOM slice in the series for metadata
    - pixel_arrays: 3D Numpy array of pixel data
    - patient_dir: Path to patient root directory
    - landmark_name: Name of landmark in JSON file
    - crop_size_cm: Crop size in centimeters [optional] (default: (5.0,
    5.0, 2.5))
    - voxel_spacing: Voxel spacing in mm (x, y, z) [optional]
    - offset_cm: Offset from landmark in centimeters (x, y, z) [optional]
    - ear_side: Side of ear to crop around (left or right) [optional]
    """

    if patient_dir is None or landmark_name is None:
```

```
32          raise ValueError("Patient directory and name must be provided for
    cropping.")
33
34      patient_id = patient_dir.split("\\")[-1]
35      landmark_path = os.path.join(
36          patient_dir, "_desc", f"{patient_id}-landmarks-ras.json"
37      )
38
39      try:
40          with open(landmark_path, mode="r", encoding="utf-8") as f:
41              landmark_json = json.load(f)
42              landmark = landmark_json["landmarks"][landmark_name]
43      except FileNotFoundError as exc:
44          raise FileNotFoundError(f"Landmark file not found at {
    landmark_path}") from exc
45      except json.JSONDecodeError as exc:
46          raise ValueError(f"Error decoding JSON file at {landmark_path}")
    from exc
47      except KeyError as exc:
48          raise KeyError(f"Landmark {landmark_name} not found in JSON file")
     from exc
49
50      if voxel_spacing is None:
51          voxel_spacing = (ds.PixelSpacing[1], ds.PixelSpacing[0], ds.
    SliceThickness)
52
53      # Apply offset based on ear side
54      if offset_cm is not None and ear_side is not None:
55          if ear_side == "left":
56              offset_cm = [crop_size_cm[0] / 5, crop_size_cm[1] / 4, 0]
57          elif ear_side == "right":
58              offset_cm = [-crop_size_cm[0] / 5, crop_size_cm[1] / 4, 0]
59
60      # Get affine matrix and convert landmark coordinates from RAS space
    to voxel space
61      if voxel_spacing is None:
62          voxel_spacing = (ds.PixelSpacing[1], ds.PixelSpacing[0], ds.
    SliceThickness)
63      affine_matrix = _get_affine_matrix(ds, voxel_spacing)
```

```
    [center_x, center_y, center_z] = _ras_2_vox(landmark, affine_matrix)

    # Apply offset if provided
    if offset_cm is not None:
        offset_mm = [o * 10 for o in offset_cm]  # Convert offset to mm
        offset_vox = [
            int(offset_mm[0] / voxel_spacing[0]),
            int(offset_mm[1] / voxel_spacing[1]),
            int(offset_mm[2] / voxel_spacing[2]),
        ]
        center_x += offset_vox[0]
        center_y += offset_vox[1]
        center_z += offset_vox[2]

    crop_size_mm = [c * 10 for c in crop_size_cm]

    # Convert crop sizes to pixels/slices
    crop_size_x = int(crop_size_mm[0] / voxel_spacing[0])  # X-dimension
    crop_size_y = int(crop_size_mm[1] / voxel_spacing[1])  # Y-dimension
    crop_size_z = int(crop_size_mm[2] / voxel_spacing[2])  # Z-dimension

    # Calculate crop bounds
    x_start = max(center_x - crop_size_x // 2, 0)
    x_end = min(center_x + crop_size_x // 2, pixel_arrays.shape[2])
    y_start = max(center_y - crop_size_y // 2, 0)
    y_end = min(center_y + crop_size_y // 2, pixel_arrays.shape[1])
    z_start = max(center_z - crop_size_z // 2, 0)
    z_end = min(center_z + crop_size_z // 2, pixel_arrays.shape[0])

    # Crop the 3D array
    cropped_volume = pixel_arrays[z_start:z_end, y_start:y_end, x_start:
    x_end]

    return cropped_volume


def _get_affine_matrix(
    ds: dicom.dataset.FileDataset, voxel_spacing: tuple = None
) -> np.array:
```

```
    """
    Calculate affine matrix from DICOM slice.

    Args:
    - ds: First DICOM slice in the series for metadata
    - voxel_spacing: Voxel spacing in mm (x, y, z) [optional]
    """

    if voxel_spacing is None:
        voxel_spacing = (ds.PixelSpacing[1], ds.PixelSpacing[0], ds.SliceThickness)

    # Extract the relevant metadata from the DICOM dataset
    image_position_patient = np.array(ds.ImagePositionPatient)  # (x0, y0, z0)
    image_orientation_patient = np.array(ds.ImageOrientationPatient)  # (x_dir, y_dir)
    pixel_spacing = np.array([voxel_spacing[0], voxel_spacing[1]])  # (dx, dy)
    slice_thickness = float(voxel_spacing[2])  # dz

    # Construct the direction vectors
    x_dir = image_orientation_patient[:3]
    y_dir = image_orientation_patient[3:6]
    slice_dir = np.cross(x_dir, y_dir)  # Compute the slice direction (z direction)

    # Build the affine transformation matrix
    dx, dy = pixel_spacing
    dz = slice_thickness
    affine_matrix = np.array(
        [
            [
                x_dir[0] * dx,
                y_dir[0] * dy,
                slice_dir[0] * dz,
                image_position_patient[0],
            ],
            [
```

```
                x_dir[1] * dx,
                y_dir[1] * dy,
                slice_dir[1] * dz,
                image_position_patient[1],
            ],
            [
                x_dir[2] * dx,
                y_dir[2] * dy,
                slice_dir[2] * dz,
                image_position_patient[2],
            ],
            [0, 0, 0, 1],
        ]
    )

    return affine_matrix


def _ras_2_vox(ras_coords: np.array, affine_matrix: np.array) -> np.array:
    """
    Convert RAS coordinates to voxel coordinates.

    Args:
    - ras_coords: Coordinates in RAS space
    - affine_matrix: Affine matrix calculated using _get_affine_matrix
    """

    # Invert affine matrix
    affine_matrix = np.linalg.inv(affine_matrix)

    # Add homogeneous coordinate
    ras_coords = np.append(ras_coords, 1)

    # Flip signs for x and y coordinates to match LPS orientation
    ras_coords[0] *= -1
    ras_coords[1] *= -1

    # Compute voxel coordinates, take first 3 elements
```

```
    voxel_coords = np.dot(affine_matrix, ras_coords)[:3]

    return np.round(voxel_coords).astype(int)
```

Listing B.1: Code for cropping the CT data around a given landmark.

```
from torch import nn

class CustomModel(nn.Module):
    def __init__(self, task="regression"):
        super().__init__()
        self.task = task
        self.output_dim = 1 if task == "regression" else 2

    def predict(self, ct_data, imp_mat, cochlear_shape):
        self.eval()
        if self.task == "classification":
            return np.max(
                self.forward(ct_data, imp_mat, cochlear_shape).detach().
    numpy(), axis=1
            )

        return self.forward(ct_data, imp_mat, cochlear_shape)

    def forward(self, ct_data, imp_mat, cochlear_shape):
        raise NotImplementedError

    def count_parameters(self):
        return sum(p.numel() for p in self.parameters() if p.
    requires_grad)

    def name(self):
        return self.__class__.__name__


class MatrixOnly(CustomModel):
    def __init__(self, layers=None, task="regression"):
        super().__init__(task=task)
        if layers is None:
            layers = [144, 64, 32, 16, self.output_dim]
```

```
        layers_list = []
        for i in range(len(layers) - 1):
            layers_list.append(nn.Linear(layers[i], layers[i + 1]))
            if i < len(layers) - 2:  # No ReLU after the last layer
                layers_list.append(nn.ReLU())
        self.network = nn.Sequential(*layers_list)

    def forward(self, ct_data, imp_mat, cochlear_shape):
        imp_mat = imp_mat.view(-1, 12 * 12)
        return self.network(imp_mat)
```

Listing B.2: Code for the implementation of MLP based impedance matrix only models.

```
import torch
from torch import nn
from torchvision.models import (
    inception_v3,
    Inception_V3_Weights,
)

class Attention(nn.Module):
    def __init__(self, hidden_dim):
        super().__init__()
        self.attention_weights = nn.Linear(hidden_dim, 1, bias=False)

    def forward(self, lstm_out):
        scores = self.attention_weights(lstm_out)
        scores = torch.softmax(scores, dim=1)
        context = (scores * lstm_out).sum(1)
        return context, scores

class InceptionTransfer(CustomModel):

    def __init__(
        self, task="regression", hidden_dim=512, num_layers=1,
    unfreeze_layer_names=None
    ):
        super().__init__(task=task)
        inception = inception_v3(weights=Inception_V3_Weights.
    IMAGENET1K_V1)
```

```
        inception.fc = Identity()
        self.feature_extractor = inception
        for param in self.feature_extractor.parameters():
            param.requires_grad = False

        self.lstm = nn.LSTM(
            2048,
            hidden_dim,
            num_layers=num_layers,
            batch_first=True,
            bidirectional=True,
        )
        self.attention = Attention(hidden_dim * 2)
        self.final_layer = nn.Sequential(
            nn.Linear(hidden_dim * 2 + 144 + 2, 128),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(128, 64),
            nn.ReLU(),
            nn.Dropout(0.3),
            nn.Linear(64, self.output_dim),
        )
        if unfreeze_layer_names is not None:
            unfreeze_layers(self, unfreeze_layer_names)

    def forward(self, ct_data, imp_mat, cochlear_shape):
        batch_size, channels, num_slices, _, _ = ct_data.size()
        slice_features = []
        for i in range(num_slices):
            ct_slice = ct_data[:, :, i, :, :]
            ct_slice = ct_slice.repeat(
                1, 3, 1, 1
            )  # Repeat channel dimension to match InceptionNet input
            slice_feature = self.feature_extractor(ct_slice)
            if self.training:
                slice_features.append(
                    slice_feature[0]
                )  # Use only the first output (second output is
    auxiliary)
```

```
            else:
                slice_features.append(slice_feature)

        slice_features = torch.stack(slice_features, dim=1)
        lstm_out, _ = self.lstm(slice_features)
        context, _ = self.attention(lstm_out)
        imp_mat = imp_mat.view(-1, 12 * 12)
        all_features = torch.cat((context, imp_mat, cochlear_shape), dim
    =1)
        output = self.final_layer(all_features)
        return output

def unfreeze_layers(model, layer_names):
    for layer_name in layer_names:
        layers = get_layer_parameters(model, layer_name)
        for layer in layers:
            layer.requires_grad = True


def get_layer_parameters(model, layer_name):
    params = []
    for name, param in model.named_parameters():
        if layer_name in name:
            params.append(param)
    return params
```

Listing B.3: Code for the implementation of the Inception-v3 Transfer model.

C. Full List of Cases

Table C.1.: Overview of the whole dataset used for this study. The validation fold is whichever fold the case appeared on the validation dataset for 10-fold cross-validation. CB: Cochlear Base; LID: Linear Insertion Depth; ExtCoch: Extracochlear; Val: Validation.

N	ID	SIDE	CB Shape (mm)		LID (mm)	ExtCoch.	Val. Fold
			Length	Width			
1	I02b	RIGHT	8.85	6.23	3.01	-	1
2	I08b	RIGHT	9.82	6.80	1.84	-	1
3	I09a	LEFT	9.03	7.08	1.97	-	1
4	I182a	LEFT	9.14	7.49	1.72	-	1
5	I21a	LEFT	9.47	7.51	1.80	-	1
6	I30b	RIGHT	9.57	6.88	1.89	-	1
7	I318a	LEFT	8.67	6.53	0.55	-	1
8	I374a	LEFT	8.49	6.12	-4.68	2	1
9	I40	RIGHT	8.44	6.65	2.81	-	1
10	I60	RIGHT	8.79	7.17	-3.75	2	1
11	I63a	LEFT	8.80	7.13	1.78	-	1
12	I67a	LEFT	9.27	7.04	-1.71	1	1
13	P03b	RIGHT	9.64	7.72	0.79	-	1
14	I149b	RIGHT	9.35	7.23	3.23	-	1
15	I157b	RIGHT	9.49	7.18	2.58	-	1
16	B017b	RIGHT	9.10	7.00	-0.02	1	2
17	Bim307a	LEFT	9.62	7.16	2.97	-	2

Continued on next page

Continued from previous page

N	ID	SIDE	CB Shape (mm)		LID (mm)	ExtCoch.	Val. Fold
			Length	Width			
18	I01b	RIGHT	8.71	6.91	1.16	-	2
19	I112b	RIGHT	9.21	7.27	1.99	-	2
20	I117b	RIGHT	8.27	6.85	0.44	-	2
21	I153a	LEFT	9.32	7.20	3.81	-	2
22	I19	RIGHT	9.64	7.31	1.41	-	2
23	I30a	LEFT	9.23	6.83	-0.75	1	2
24	I322a	LEFT	8.46	6.14	2.80	-	2
25	I34	LEFT	9.51	7.29	2.03	-	2
26	I89b	RIGHT	9.48	7.30	-4.04	2	2
27	I98a	LEFT	9.07	7.06	1.69	-	2
28	I01a	LEFT	8.69	6.76	2.41	-	2
29	I128b	RIGHT	9.68	7.88	1.36	-	2
30	I130a	LEFT	8.85	6.81	2.35	-	2
31	B008a	LEFT	8.56	6.59	1.96	-	3
32	C03a	LEFT	8.93	6.25	2.87	-	3
33	I05	RIGHT	8.00	6.23	-0.82	1	3
34	I12b	RIGHT	9.12	6.74	1.58	-	3
35	I13a	LEFT	9.55	7.08	2.56	-	3
36	I148b	RIGHT	9.66	7.69	2.34	-	3
37	I24b	RIGHT	9.48	6.65	2.54	-	3
38	I46a	LEFT	9.38	7.01	2.08	-	3
39	I50	LEFT	9.03	6.94	2.71	-	3
40	I62	RIGHT	8.59	6.30	-1.08	1	3
41	I64a	LEFT	8.48	6.50	2.36	-	3
42	I90b	RIGHT	10.03	8.21	1.22	-	3
43	P05a	LEFT	9.57	6.60	2.50	-	3
44	I82a	LEFT	8.82	7.07	2.49	-	3

Continued on next page

Continued from previous page

N	ID	SIDE	CB Shape (mm)		LID (mm)	ExtCoch.	Val. Fold
			Length	Width			
45	I04b	RIGHT	9.79	7.26	3.19	-	4
46	I109a	LEFT	9.62	7.39	4.05	-	4
47	I10b	RIGHT	9.03	6.59	2.25	-	4
48	I11	RIGHT	9.29	6.63	2.61	-	4
49	I111b	RIGHT	8.90	6.69	1.50	-	4
50	I126b	RIGHT	9.67	7.93	4.02	-	4
51	I140b	RIGHT	10.10	7.86	3.82	-	4
52	I142a	LEFT	8.48	7.21	3.84	-	4
53	I151b	RIGHT	9.31	6.98	0.80	-	4
54	I318b	RIGHT	9.82	6.51	2.51	-	4
55	I31a	LEFT	8.73	6.57	1.93	-	4
56	I388a	LEFT	8.21	6.41	-1.29	1	4
57	I48	RIGHT	9.12	7.07	2.88	-	4
58	I61	RIGHT	8.85	6.87	-2.87	2	4
59	I12a	LEFT	9.24	6.62	1.80	-	5
60	I134b	RIGHT	9.52	7.68	2.29	-	5
61	I159a	LEFT	9.44	7.29	2.94	-	5
62	I183b	RIGHT	9.72	7.99	3.49	-	5
63	I26b	RIGHT	9.55	6.73	2.29	-	5
64	I36	RIGHT	9.25	6.77	2.85	-	5
65	I390b	RIGHT	9.16	7.32	0.91	-	5
66	I70b	RIGHT	8.67	7.08	3.77	-	5
67	I77b	RIGHT	9.28	7.12	2.90	-	5
68	I86b	RIGHT	8.32	6.43	1.95	-	5
69	I135a	LEFT	9.55	7.60	6.31	-	5
70	I49	RIGHT	9.53	7.48	0.42	-	5
71	I69a	LEFT	9.08	7.38	1.50	-	5

Continued on next page

Continued from previous page

N	ID	SIDE	CB Shape (mm)		LID (mm)	ExtCoch.	Val. Fold
			Length	Width			
72	I73b	RIGHT	9.53	6.76	0.40	-	5
73	B019b	RIGHT	8.51	6.73	1.89	-	6
74	Bim304a	LEFT	9.76	6.95	3.84	-	6
75	I135b	RIGHT	9.61	7.49	2.69	-	6
76	I138b	RIGHT	8.89	6.71	-0.77	1	6
77	I150a	LEFT	9.12	6.69	-0.15	1	6
78	I16a	LEFT	8.84	6.53	-0.80	1	6
79	I324b	RIGHT	9.47	6.36	-1.74	1	6
80	I46b	RIGHT	9.45	7.42	1.83	-	6
81	I47	RIGHT	9.94	7.42	3.63	-	6
82	I55a	LEFT	9.16	7.78	-0.51	1	6
83	I92a	LEFT	9.32	7.02	-0.33	1	6
84	I118b	RIGHT	9.08	6.76	1.07	-	6
85	I137b	RIGHT	8.75	6.73	2.65	-	6
86	I58	RIGHT	8.62	6.81	0.56	-	6
87	B013a	RIGHT	8.24	5.74	4.24	-	7
88	B017a	LEFT	10.27	6.79	2.00	-	7
89	Bim303a	LEFT	8.39	5.74	3.61	-	7
90	Bim305a	LEFT	9.49	7.36	4.63	-	7
91	I03b	RIGHT	8.60	6.79	0.12	-	7
92	I08a	LEFT	9.25	7.18	2.63	-	7
93	I14	LEFT	8.75	7.07	1.05	-	7
94	I18b	RIGHT	9.17	6.53	1.56	-	7
95	I336a	LEFT	9.32	6.93	1.83	-	7
96	I70a	LEFT	8.88	6.95	1.71	-	7
97	P02b	RIGHT	9.14	6.65	2.42	-	7
98	P03a	LEFT	9.95	7.36	-1.60	1	7

Continued on next page

Continued from previous page

N	ID	SIDE	CB Shape (mm)		LID (mm)	ExtCoch.	Val. Fold
			Length	Width			
99	I09b	RIGHT	9.01	7.15	-1.19	1	7
100	I55b	RIGHT	8.77	7.58	2.29	-	7
101	I144a	LEFT	9.53	7.16	2.87	-	8
102	I158b	RIGHT	9.97	7.81	1.46	-	8
103	I15a	LEFT	9.40	6.97	1.98	-	8
104	I181a	LEFT	8.85	6.64	1.10	-	8
105	I22	RIGHT	9.18	7.08	1.38	-	8
106	I26a	LEFT	9.89	6.97	2.61	-	8
107	I39	LEFT	9.26	7.26	2.82	-	8
108	I53	RIGHT	8.24	6.75	1.50	-	8
109	I56	RIGHT	8.80	6.38	0.76	-	8
110	I57a	LEFT	8.57	6.74	1.34	-	8
111	I75a	LEFT	9.70	7.14	4.06	-	8
112	I100b	RIGHT	8.67	6.82	-1.21	1	8
113	I147b	RIGHT	9.30	6.46	1.70	-	8
114	I75b	RIGHT	9.55	7.34	2.31	-	8
115	B008b	RIGHT	8.42	6.11	-1.04	1	9
116	I105a	LEFT	9.20	6.92	2.68	-	9
117	I129a	LEFT	9.59	7.35	2.36	-	9
118	I132a	LEFT	8.90	6.91	2.01	-	9
119	I13b	RIGHT	9.45	7.30	2.52	-	9
120	I21b	RIGHT	9.43	7.40	1.45	-	9
121	I24a	LEFT	9.32	7.04	1.14	-	9
122	I28a	LEFT	8.60	6.42	1.88	-	9
123	I35a	LEFT	9.45	7.20	3.77	-	9
124	I85a	LEFT	9.26	7.51	2.80	-	9
125	I131a	LEFT	8.38	6.68	2.07	-	9

Continued on next page

Continued from previous page

N	ID	SIDE	CB Shape (mm)		LID (mm)	ExtCoch.	Val. Fold
			Length	Width			
126	I136b	RIGHT	9.20	7.35	-1.46	1	9
127	I17b	RIGHT	8.58	6.61	-3.28	2	9
128	I63b	RIGHT	8.52	7.10	0.87	-	9
129	I04a	LEFT	9.87	7.45	3.91	-	10
130	I102b	RIGHT	9.88	7.35	2.02	-	10
131	I183a	LEFT	9.69	8.08	3.26	-	10
132	I327b	RIGHT	9.46	6.49	1.07	-	10
133	I341b	RIGHT	8.85	6.54	4.43	-	10
134	I52	LEFT	8.63	6.64	0.66	-	10
135	I57b	RIGHT	8.68	6.72	2.35	-	10
136	I68b	RIGHT	8.71	6.40	0.85	-	10
137	I74a	LEFT	9.40	7.51	4.91	-	10
138	I74b	RIGHT	9.23	7.40	2.25	-	10
139	I80a	LEFT	9.11	7.51	-0.02	1	10
140	P04a	LEFT	8.54	6.71	2.27	-	10
141	I111a	LEFT	9.22	6.93	1.85	-	10
142	I87b	RIGHT	8.92	6.53	1.59	-	10

Abbreviations

CI Cochlear Implant

CT Computed Tomography

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

LID Linear Insertion Depth

MRI Magnetic Resonance Imaging

SNHL Sensorineural Hearing Loss

IFT Impedance Field Telemetry

RMSE Root Mean Squared Error

MAE Mean Absolute Error

DICOM Digital Imaging and Communications in Medicine

NIfTI Nueuroimaging Informatics Technology Initiative

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

CT-FM Computed Tomography-Foundation Model

MLP Multilayer Perceptron

MM Mean Model

List of Figures

List of Tables

List of Listings

Bibliography

[1] C. M. Conway *et al.*, "The Importance of Sound for Cognitive Sequencing Abilities: The Auditory Scaffolding Hypothesis," en, *Current Directions in Psychological Science*, vol. 18, no. 5, pp. 275–279, Oct. 2009, ISSN: 0963-7214, 1467-8721. DOI: 10.1111/j.1467-8721.2009.01651.x.

[2] *World Report on Hearing*, eng, 1st ed. Geneva: World Health Organization, 2021, ISBN: 9789240020481.

[3] A. Shukla *et al.*, "Hearing Loss, Loneliness, and Social Isolation: A Systematic Review," en, *Otolaryngology–Head and Neck Surgery*, vol. 162, no. 5, pp. 622–633, May 2020, ISSN: 0194-5998, 1097-6817. DOI: 10.1177/0194599820910377.

[4] J. B. Nadol, "Hearing Loss," en, *New England Journal of Medicine*, vol. 329, no. 15, pp. 1092–1102, Oct. 1993, ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJM199310073291507.

[5] A. Hearing, *Types of hearing loss*, 2005.

[6] T. Zahnert, "The Differential Diagnosis of Hearing Loss," *Deutsches Ärzteblatt international*, Jun. 2011, ISSN: 1866-0452. DOI: 10.3238/arztebl.2011.0433.

[7] J. E. Isaacson and N. M. Vora, "Differential diagnosis and treatment of hearing loss," *American family physician*, vol. 68, no. 6, pp. 1125–1132, 2003.

[8] R. J. Smith *et al.*, "Sensorineural hearing loss in children," en, *The Lancet*, vol. 365, no. 9462, pp. 879–890, Mar. 2005, ISSN: 01406736. DOI: 10.1016/S0140-6736(05)71047-3.

[9] X. Meng *et al.*, "Covid-19 and sudden sensorineural hearing loss: A systematic review," *Frontiers in Neurology*, vol. 13, p. 883749, 2022. DOI: 10.3389/fneur.2022.883749.

[10] A. G. Bittencourt *et al.*, "Prelingual deafness: Benefits from cochlear implants versus conventional hearing aids," en, *International Archives of Otorhinolaryngology*, vol. 16, no. 03, pp. 387–390, Jul. 2012, ISSN: 1809-9777, 1809-4864. DOI: 10.7162/S1809-97772012000300014.

[11] B. S. Wilson and M. F. Dorman, "Cochlear implants: Current designs and future possibilities," *J Rehabil Res Dev*, vol. 45, no. 5, pp. 695–730, 2008. DOI: 10.1682/jrrd.2007.10.0173.

[12] T. Vogl *et al.*, "Pre-, intra-and post-operative imaging of cochlear implants," in *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, © Georg Thieme Verlag KG, vol. 187, 2015, pp. 980–989. DOI: 10.1055/s-0035-1553413.

[13] B. Vaerenberg *et al.*, "Cochlear implant programming: A global survey on the state of the art," *The scientific world journal*, vol. 2014, no. 1, p. 501 738, 2014. DOI: 10.1155/2014/501738.

[14] R. I. Banda González *et al.*, "Fitting parameters for cochlear implant," *Boletín Médico Del Hospital Infantil de México (English Edition)*, vol. 74, no. 1, pp. 65–69, Jan. 2017, ISSN: 2444-3409. DOI: 10.1016/j.bmhime.2017.11.016.

[15] A. Kurz *et al.*, "Using anatomy-based fitting to reduce frequency-to-place mismatch in experienced bilateral cochlear implant users: A promising concept," *Journal of Personalized Medicine*, vol. 13, no. 7, p. 1109, Jul. 2023, ISSN: 2075-4426. DOI: 10.3390/jpm13071109.

[16] A. Berrington de Gonzalez *et al.*, "Epidemiological studies of ct scans and cancer risk: The state of the science," *British Journal of Radiology*, vol. 94, no. 1126, p. 20 210 471, Sep. 2021, ISSN: 0007-1285. DOI: 10.1259/bjr.20210471.

[17] M.-O. Bernier *et al.*, "Potential cancer risk associated with ct scans: Review of epidemiological studies and ongoing studies," *Progress in Nuclear Energy*, vol. 84, pp. 116–119, 2015, EUROSAFE 2013, ISSN: 0149-1970. DOI: https://doi.org/10.1016/j.pnucene.2014.07.011.

[18] C.-F. Cao *et al.*, "Ct scans and cancer risks: A systematic review and dose-response meta-analysis," *BMC cancer*, vol. 22, no. 1, p. 1238, 2022.

[19] H. E. Rice *et al.*, "Review of radiation risks from computed tomography: Essentials for the pediatric surgeon," *Journal of pediatric surgery*, vol. 42, no. 4, pp. 603–607, 2007. DOI: 10.1016/j.jpedsurg.2006.12.009.

[20] F. Wang *et al.*, "Deep learning in medicine—promise, progress, and challenges," *JAMA Internal Medicine*, vol. 179, no. 3, p. 293, Mar. 2019, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2018.7117.

[21] L. Heiliger *et al.*, "Beyond medical imaging - a review of multimodal deep learning in radiology," Feb. 2022. DOI: 10.36227/techrxiv.19103432.v1.

[22] F. Behrad and M. Saniee Abadeh, "An overview of deep learning methods for multimodal medical data mining," *Expert Systems with Applications*, vol. 200, p. 117 006, Aug. 2022, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.117006.

[23] S. Schraivogel *et al.*, "Predictive Models for Radiation-Free Localization of Cochlear Implants' Most Basal Electrode using Impedance Telemetry," *IEEE Transactions on Biomedical Engineering*, pp. 1–12, 2024, ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2024.3509527.

[24] T. L. Bruns *et al.*, "Real-Time Localization of Cochlear-Implant Electrode Arrays Using Bipolar Impedance Sensing," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 718–724, Feb. 2022, ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2021.3104104.

[25] P. Aebischer *et al.*, "Intraoperative Impedance-Based Estimation of Cochlear Implant Electrode Array Insertion Depth," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 2, pp. 545–555, Feb. 2021, ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2020.3006934.

[26] Y. Dong *et al.*, "Detection of translocation of cochlear implant electrode arrays by intracochlear impedance measurements," *Ear and Hearing*, vol. 42, no. 5, pp. 1397–1404, Apr. 2021, ISSN: 1538-4667. DOI: 10.1097/aud.0000000000001033.

[27] C. K. Giardina *et al.*, "Impedance measures during in vitro cochlear implantation predict array positioning," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 2, pp. 327–335, Feb. 2018, ISSN: 1558-2531. DOI: 10.1109/tbme.2017.2764881.

[28] L. Sijgers *et al.*, "Predicting cochlear implant electrode placement using monopolar, three-point and four-point impedance measurements," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2533–2544, Aug. 2022, ISSN: 1558-2531. DOI: 10.1109/tbme.2022.3150239.

[29] S. Schraivogel *et al.*, "Postoperative impedance-based estimation of cochlear implant electrode insertion depth," *Ear and Hearing*, vol. 44, no. 6, pp. 1379–1388, May 2023, ISSN: 1538-4667. DOI: 10.1097/aud.0000000000001379.

[30] L. Zhang *et al.*, "Transimpedance matrix can be used to estimate electrode positions intraoperatively and to monitor their positional changes postoperatively in cochlear implant patients," *Otology and Neurotology*, vol. 45, no. 4, e289–e296, Feb. 2024, ISSN: 1531-7129. DOI: 10.1097/mao.0000000000004145.

[31] A. A. Eshraghi *et al.*, "The cochlear implant: Historical aspects and future prospects," en, *Anat. Rec. (Hoboken)*, vol. 295, no. 11, pp. 1967–1980, Nov. 2012. DOI: 10.1002/ar.22580.

[32] J. M. Goins and G. Mankekar, "William f. house: The father of neurotology," en, *Cureus*, vol. 16, no. 9, e69724, Sep. 2024. DOI: 10.7759/cureus.69724.

[33] R. J. Fretz and R. P. Fravel, "Design and function: A physical and electrical description of the 3M house cochlear implant system," *Ear and Hearing*, vol. 6, no. 3, 14S–19S, 1985.

[34] F.-G. Zeng, "Celebrating the one millionth cochlear implant," en, *JASA Express Letters*, vol. 2, no. 7, p. 077 201, Jul. 2022, ISSN: 2691-1191. DOI: 10.1121/10.0012825.

[35] E. Tzvi-Minker and A. Keck, "How can we compare cochlear implant systems across manufacturers? a scoping review of recent literature," *Audiology Research*, vol. 13, no. 5, pp. 753–766, Oct. 2023, ISSN: 2039-4349. DOI: 10.3390/audiolres13050067.

[36] S. of Blausen Medical, "Medical gallery of blausen medical 2014," *WikiJournal of Medicine*, vol. 1, no. 2, 2014, ISSN: 2002-4436. DOI: 10.15347/wjm/2014.010.

[37] W. A. Kalender, "X-ray computed tomography," *Physics in Medicine and Biology*, vol. 51, no. 13, R29–R43, Jun. 2006, ISSN: 1361-6560. DOI: 10.1088/0031-9155/51/13/r03.

[38] W. C. Scarfe and A. G. Farman, "What is cone-beam ct and how does it work?" *Dental Clinics of North America*, vol. 52, no. 4, pp. 707–730, Oct. 2008, ISSN: 0011-8532. DOI: 10.1016/j.cden.2008.05.005.

[39] J. W. Casselman *et al.*, "Cone beam ct: Non-dental applications," *Journal of the Belgian Society of Radiology*, vol. 96, no. 6, p. 333, Nov. 2013, ISSN: 1780-2393. DOI: 10.5334/jbr-btr.453.

[40] J. Benson *et al.*, "A new frontier in temporal bone imaging: Photon-counting detector ct demonstrates superior visualization of critical anatomic structures at reduced radiation dose," *American Journal of Neuroradiology*, vol. 43, no. 4, pp. 579–584, Mar. 2022, ISSN: 1936-959X. DOI: 10.3174/ajnr.a7452.

[41] H. R. Harnsberger *et al.*, "Cochlear implant candidates: Assessment with ct and mr imaging.," *Radiology*, vol. 164, no. 1, pp. 53–57, Jul. 1987, ISSN: 1527-1315. DOI: 10.1148/radiology.164.1.3108956.

[42] M. H. Alam-Eldeen *et al.*, "Radiological requirements for surgical planning in cochlear implant candidates," *Indian Journal of Radiology and Imaging*, vol. 27, no. 03, pp. 274–281, Jul. 2017, ISSN: 1998-3808. DOI: 10.4103/ijri.ijri_55_17.

[43] S. S. Connell *et al.*, "Electrode migration after cochlear implantation," *Otology and Neurotology*, vol. 29, no. 2, pp. 156–159, Feb. 2008, ISSN: 1531-7129. DOI: 10.1097/mao.0b013e318157f80b.

[44] J. T. Holder *et al.*, "Prevalence of extracochlear electrodes: Computerized tomography scans, cochlear implant maps, and operative reports," *Otology and Neurotology*, vol. 39, no. 5, e325–e331, Jun. 2018, ISSN: 1537-4505. DOI: 10.1097/mao.0000000000001818.

[45] J. Bidgood W. Dean *et al.*, "Understanding and using dicom, the data interchange standard for biomedical imaging," *Journal of the American Medical Informatics Association*, vol. 4, no. 3, pp. 199–212, May 1997, ISSN: 1067-5027. DOI: 10.1136/jamia.1997.0040199.

[46] X. Li *et al.*, "The first step for neuroimaging data analysis: Dicom to nifti conversion," *Journal of Neuroscience Methods*, vol. 264, pp. 47–56, May 2016, ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2016.03.001.

[47] D. Richfield and M. Häggström, *Human anatomy planes, labeled*. Oct. 2014.

[48] T. D. DenOtter and J. Schubert, *Hounsfield Unit*. StatPearls Publishing, Treasure Island (FL), 2023.

[49] Amisha *et al.*, "Overview of artificial intelligence in medicine," *Journal of Family Medicine and Primary Care*, vol. 8, no. 7, p. 2328, 2019, ISSN: 2249-4863. DOI: 10.4103/jfmpc.jfmpc_440_19.

[50] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, S36–S40, Apr. 2017, ISSN: 0026-0495. DOI: 10.1016/j.metabol.2017.01.011.

[51] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005.

[52] R. Yamashita *et al.*, "Convolutional neural networks: An overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Jun. 2018, ISSN: 1869-4101. DOI: 10.1007/s13244-018-0639-9.

[53] A. Barragán-Montero *et al.*, "Artificial intelligence and machine learning for medical imaging: A technology review," *Physica Medica*, vol. 83, pp. 242–256, Mar. 2021, ISSN: 1120-1797. DOI: 10.1016/j.ejmp.2021.04.016.

[54] J. van der Laak *et al.*, "Deep learning in histopathology: The path to the clinic," *Nature Medicine*, vol. 27, no. 5, pp. 775–784, May 2021, ISSN: 1546-170X. DOI: 10.1038/s41591-021-01343-4.

[55] C. Li *et al.*, "Cochlear ct image segmentation based on u-net neural network," *Journal of Radiation Research and Applied Sciences*, vol. 16, no. 2, p. 100 560, Jun. 2023, ISSN: 1687-8507. DOI: 10.1016/j.jrras.2023.100560.

[56] F. Heutink *et al.*, "Multi-scale deep learning framework for cochlea localization, segmentation and analysis on clinical ultra-high-resolution ct images," *Computer Methods and Programs in Biomedicine*, vol. 191, p. 105 387, Jul. 2020, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2020.105387.

[57] Y. Fan *et al.*, "A unified deep-learning-based framework for cochlear implant electrode array localization," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, 2023, pp. 376–385, ISBN: 9783031439964. DOI: 10.1007/978-3-031-43996-4_36.

[58] Y. Chi *et al.*, "A deep-learning-based method for the localization of cochlear implant electrodes in ct images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, Apr. 2019. DOI: 10.1109/isbi.2019.8759536.

[59] A. Althnian *et al.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Applied Sciences*, vol. 11, no. 2, p. 796, Jan. 2021, ISSN: 2076-3417. DOI: 10.3390/app11020796.

[60] J. Ngiam *et al.*, "Multimodal deep learning," *In ICML*, vol. 11, pp. 689–696, 2011.

[61] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, Nov. 2017, ISSN: 1053-5888. DOI: 10.1109/msp.2017.2738401.

[62] C. Cui *et al.*, "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review," *Progress in Biomedical Engineering*, vol. 5, no. 2, p. 022 001, Apr. 2023, ISSN: 2516-1091. DOI: 10.1088/2516-1091/acc2fe.

[63] S.-C. Huang *et al.*, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, no. 1, Oct. 2020, ISSN: 2398-6352. DOI: 10.1038/s41746-020-00341-z.

[64] S. R. Stahlschmidt *et al.*, "Multimodal deep learning for biomedical data fusion: A review," *Briefings in Bioinformatics*, vol. 23, no. 2, Jan. 2022, ISSN: 1477-4054. DOI: 10.1093/bib/bbab569.

[65] A. Fedorov *et al.*, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012, ISSN: 0730-725X. DOI: 10.1016/j.mri.2012.05.001.

[66] T. Khurayzi *et al.*, "Direct measurement of cochlear parameters for automatic calculation of the cochlear duct length," *Annals of Saudi Medicine*, vol. 40, no. 3, pp. 212–218, May 2020, ISSN: 0975-4466. DOI: 10.5144/0256-4947.2020.218.

[67] M. J. Cardoso *et al.*, *Monai: An open-source framework for deep learning in healthcare*, 2022. DOI: 10.48550/ARXIV.2211.02701.

[68] S. Wiki, *Coordinate systems — slicer wiki*, [Online; accessed 24-February-2025], 2023.

[69] S. Pai *et al.*, *Vision foundation models for computed tomography*, 2025. DOI: 10.48550/ARXIV.2501.09001.

[70] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, ISSN: 1530-888X. DOI: 10.1162/neco.1989.1.4.541.

[71] C. Szegedy *et al.*, *Going deeper with convolutions*, 2014. DOI: 10.48550/ARXIV.1409.4842.

[72] C. Szegedy *et al.*, *Rethinking the inception architecture for computer vision*, 2015. DOI: 10.48550/ARXIV.1512.00567.

[73] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[74] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, ISSN: 1530-888X. DOI: 10.1162/neco.1997.9.8.1735.

[75] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. DOI: 10.48550/ARXIV.1412.6980.

[76] J. Lahoud *et al.*, *3d vision with transformers: A survey*, 2022. DOI: 10.48550/ARXIV.2208.04309.