

University of Pecs
Faculty of Engineering and Information Technology
Biomedical Engineering MSc Program

DIPLOMA THESIS

A Novel Radiological Software for Automatically Detecting the Inner Ear and Classifying Normal From Malformed Anatomy



**BIOMEDICAL
ENGINEERING
PROGRAMME**

Author:

Abdulrahman Alkojak Almasri

Supervisors:

Zsolt Kisander (MSc, PhD)

Andras Kedves (MSc, PhD)

Pecs
2024

UNIVERSITY OF PECS
BIOMEDICAL ENGINEERING MASTER PROGRAM
MSC THESIS TASK DESCRIPTION

Abdulrahman Alkojak Almasri

ICA06Y, Number:/2024

Candidate for MSc degree in Biomedical Engineering

In the Research and Development Department at MED-EL (Electrode and Inner Ear Malformation Research Team), a new project was initiated to create an automated workflow for extracting the inner ear region from temporal CT scans and classifying it according to the shape and the 12 types of malformations. My job as an intern was to prepare the dataset, develop an inner ear detection algorithm, build a neural network to perform normal vs. abnormal classification, develop a data pipeline to crop the inner ear region, process it, and store it automatically, and build demo software to showcase the detection and classification.

Perform the following tasks:

1. Data collection and management:

- (a) Received a hard disk containing 750 case analysis for different inner ear malformations. The cases were in 3D Slicer view format.
- (b) Manually extracting the DICOM scans for the best study possible for each patient.
- (c) Manually segmenting the inner ear structures and exporting it as STL format.
- (d) Building an Excel sheet for patient information and metadata.
- (e) Extracting and Mid-Modular section for the segmented inner ear
- (f) structures.

2. Cropping algorithm development:

- (a) Two cropping algorithms were developed, hounsfield unit based and K- means based.

-
- (b) The first algorithm was developed by Dr. Andras as a preliminary idea, then I have added modifications to it based on the input i received. Then I came up with a novel methodology based on K-Means algorithm.
 - (c) The two algorithms then were assessed, compared and evaluated in a manual manner.
3. Classification model development:
- (a) A 3D Convolutional Neural Network has been developed using the cropped volumes managed from previous steps.
 - (b) Multiple architectures have been tested to get the optimal results.
4. End-to-End data processing pipeline:
- (a) I have developed multiple python scripts for automatically loading the DICOM scans, pre-processing the volumes (orientation correction, Hounsfield unit adjustment, trimming to eliminate the background, splitting the volumes in half, etc.)
 - (b) Performing the cropping algorithm to the individual inner ear structures (left and right sides).
 - (c) Post-processing the cropped volume with the help of the patient's metadata.
 - (d) Flipping the inner ear's right side cropped volumes to the left side.
 - (e) Storing the final volumes in folders according to the malformation type.
 - (f) Manually checking the cropped volumes and eliminating the bad scans.
 - (g) Preparing the data for the training, validation and testing.
5. OTO-AI demo software development:
- (a) A demo software was developed using python and Qt5 module.
 - (b) To visualize the DICOM scans, perform the auto and manual cropping.
 - (c) And finally, to diagnose the cropped volumes accordingly.
 - (d) Different versions were developed to test different features and functionalities.
6. Participation in the European Congress of Radiology:

-
- (a) The work has also been accepted by the European Community of Radiology ECR congress.
 - (b) Abdulrahman presented the project on 28th of February during the student session.

Internal supervisor: Kisander Zsolt (Msc, PhD.), University of Pecs, Faculty of Engineering and Information Technology, Automation department, zsolt.kisander@mik.pte.hu.

External supervisor: Kedves Andras (Msc, PhD.), MED-EL Medical Electronics, R&D department, Andras.kedves@medel.com.

Proposed subject, topic related to the thesis, required for the final examination (Curriculum): Artificial Intelligence

Pecs, 18 March 20204

Stamp

Signature of the internal supervisor

Declaration

I declare, that this diploma thesis is the result of my own work. All references and external works have been identified and cited. I have not used any other external help.

The results of my diploma thesis can be used by the university for its own purposes free of charge.

Pecs, 25 May 2024

Signature of the student.

Abstract

Background To develop effective radiological software that could read Digital Imaging and Communications in Medicine (DICOM) files, automatically crop the inner ear based on head computed tomography (CT), and classify normal and abnormal inner ear anatomy (IEM) automatically.

Methods We retrospectively evaluated 2053 patients from three hospitals and extracted 1200 inner ear CTs to import, crop, and use for artificial intelligence (AI) training, testing, and validation. Automated cropping algorithms based on CTs were created to crop the inner ear volume, along with a simple graphical user interface (GUI). Using these crops as input, we created a deep learning convolutional neural network (DL CNN) with 5-fold cross-validation to determine whether the inner ear anatomy is abnormal or normal. Cochlear hypoplasia (CH), ossification, incomplete partition (IP) type I, IP III, and common cavity (data equally distributed) were selected. The cropping tool and the AI model were both verified.

Results The newly developed DICOM viewer/software can read CT files, crop a volume that contains the inner ear, and classify the anatomy as either normal or malformed. The cropping tool provided an average accuracy of 92.25%. The area under the curve (AUC) is 0.86 (95% confidence interval (CI): 0.81-0.91) on DL. Accuracy, precision, recall, and F1 scores are 0.812, 0.791, 0.8, and 0.766, respectively.

Conclusion We present the development and validation of a potentially fully automated workflow for classifying normal versus abnormal inner ear anatomy. The tool demonstrated good diagnostic accuracy for risk stratification but must be supervised by a decision-maker.

Keywords: artificial intelligence, deep learning, radiology, digital health

Abbreviations

- Area under curve (AUC).
- Three-dimensional (3D).
- Computed tomography (CT).
- Convolutional neural network (CNN).
- Confidence interval (CI).
- Deep learning (DL).
- Digital Imaging and Communications in Medicine (DICOM).
- Ear, nose, and throat (ENT).
- Hounsfield unit (HU).
- Incomplete partition type III (IPIII).
- Inner ear malformation (IEM).
- Gradient-weighted Class Activation Mapping (Grad-CAM).
- Mann-Whitney U rank test (MWU).
- Px: pixels (Px).
- Receiver operating characteristic (ROC).
- Region-Based Convolutional Neural Network (RCNN).

Highlights

- DL CNNs were trained using 1200 head CT images obtained from 3 centers.
- K-means based algorithm performance was superior to the pure HU-based (Accuracy: 87.2% and 92.25%, respectively).
- AI classifier accuracy, precision, recall, and F1 scores are 0.812, 0.791, 0.8, and 0.766, respectively.
- Performance shown by the area under the curve (AUC) is 0.86 (95% CI: 0.81-0.91).
- GUI developed for efficient feedback.

Contents

1	Introduction	12
2	Background	14
2.1	Anatomy of the ear and physiology of hearing	14
2.1.1	The outer ear	15
2.1.2	The middle ear	15
2.1.3	The inner ear	16
2.2	Inner ear malformations IEMs	18
2.3	K-Means clustering	21
2.4	Convolutional Neural Networks	22
2.4.1	Convolution layers	23
2.4.2	Pooling layers	25
2.4.3	Fully connected layers	26
2.4.4	Dropout layers	26
2.4.5	Batch normalization	26
3	Methodology and Theory	28
3.1	Patient cohorts, image analyses, pre-processing	28
3.2	Cropping algorithms	29
3.2.1	Hounsfield units and distance-based cropping	29
3.2.2	Two-staged spatial auto k-means clustering cropping	30
3.3	Dataset, model developments, settings, and evaluation	31
3.4	Validation, statistic evaluation	34
4	Results	35
4.1	Cropping algorithms	35
4.2	Data pipeline	36
4.3	3D CNN metrics	36

4.4	OTO-AI demo software	36
5	Discussion	39
6	Acknowledgments	41
7	Appendix 1	42
7.1	Supplementary methods	42
7.1.1	Institution overview	42
7.2	Supplementary references	43
7.3	Supplementary figures	43
7.4	Supplementary codes and usage	43
7.4.1	Command Line Interface usage	43
7.4.2	Inner ear localization and cropping	44

List of Figures

1	Detailed anatomical structure of the ear.	15
2	Line drawings of left and right bony labyrinths of a generalized mammal in dorsal view	17
3	Normal inner ear anatomy demonstrated on axial CT	18
4	Different inner ear malformations in the axial, oblique coronal views and 3d segmented.	21
5	3D k-means clustering results	23
6	General 3D CNN architecture	24
7	Convolution operation example	25
8	Data processing and distribution	29
9	General diagram of Hounsfield units and distance-based cropping algorithm.	30
10	General diagram of Two-staged spatial auto k-means clustering cropping algorithm. .	31
11	Three-dimensional convolutional neural network model architecture.	32
12	Training routine and data pipeline.	33
13	Kmeans cropping algorithm sample results	37
14	Grad-CAM for the activation maps	38
15	OTOai software prototype GUI	38
16	Deep learning convolutional neural network model architecture.	44
17	Full project architecture	45

List of Tables

1	Accuracy table for the proposed algorithms on the pre operative data	35
2	Accuracy comparison between algorithms on post and pre-operative scans.	36
3	Overall accuracies table for cropping algorithm between normal and abnormal anatomies	36
4	3D Convolutional Neural Network metrics	36

Chapter 1

Introduction:

According to reports from all over the world, children with congenital hearing loss have an inner ear malformation (IEM) rate of 20%-30%.[1, 2], Gross malformations are a contraindication to cochlear implantation surgery and major deformities may cause surgical issues. Since computed tomography (CT) provides important information on the state of the inner ear and other abnormalities both before and after surgery, it has been the main method of diagnosis.[3] because the inner ear is a complex anatomical structure It can be challenging for non-specialists to make an accurate diagnosis of inner ear illnesses.[4] It is difficult to identify inner ear malformations (IEMs) using radiological imaging. However, automatic detection of IEMs using deep learning (DL) could lead to the development of new methods for diagnosing all particular types of IEMs. DL is a type of artificial intelligence that can learn to recognize patterns in data, and it has been shown to be effective for automatically detecting anatomical structures in medical images.

Localization of the malformed inner ear does not exist as far as we know, however, there are attempts to diagnose it semiautomatically (with manually adjusted inner ear localization)[5] or using methods developed for normal anatomy and finding the inner ear with unknown accuracy[6, 7].

Despite the fact that DL has become more common in the medical community[8], in particular when contrasting human and machine performance[9], and DL is employed to evaluate temporal bone CTs[5], no study on fully automated inner ear crop and classification has been published before.

To analyze the complex structures/objects of the temporal bone, 3 dimensional - convolutional neural network (3D-CNN) would be required. The recent development of CNN architectures and powerful computer graphics processing units with a lot of memory has made 3D-DL useful for medical image analysis.[10]

Additionally, it was mentioned that, when compared to unsupervised techniques, supervised abnormality detection with only 64 abnormal samples was excellent. In supervised deep learning, training data with labels like "normal" or "abnormal" are used to build a standard classification CNN

that classifies test data.[11] To the best of our knowledge, a fully automatized DL workflow for inner ear malformation diagnosis has not yet been developed and validated in any of the available studies. Our AI technology can help professionals evaluate CT images more accurately. The tool might also be able to provide crucial second opinion information needed for patient-tailored diagnosis and surgery prior to or after the implantation of a cochlear implant. One study only covers non-genital inner ear malformations amongst many malformations[6] and do not have diverse (based on 1 country), and robust number of input data[7]. Another study classified inner ear with outstanding accuracy, however it provided a semiautomatic way of inner ear localization, which made it time-consuming and impossible to implement to real-life workflow and was restricted to using a single institute technique[5]. The present study focuses on software development for the fully automatized diagnosis of wide range of inner ear malformations, by evaluating the effectiveness of supervised DL on CT images and whether abnormal regions were accurately detected. We also developed an inner ear identifier method and assessed the accuracy of it. Deep learning-based categorization needs data from multi-center studies and large-scale evaluations, but none of these studies have been carried out for IEM. In order to fill this unmet need, we gathered information from three cohorts and countries, created and evaluated an inner ear cropping tool, and assessed the diagnostic performance of a deep learning-based classifier. As a result we developed an AI-based software for inner ear diagnosis, which we call OTOai.

Chapter 2

Background:

This study has extensively utilized deep learning, programming, and image processing techniques, in addition to focusing on the inner ear's anatomical structure and the various types and manifestations of malformations. The scientific foundation for this research will be thoroughly explained in this chapter.

2.1 Anatomy of the ear and physiology of hearing:

One of the primary senses, hearing, like vision, is crucial for communication and distant warning. It can be used to signal alertness, joy, or terror. It is the deliberate enjoyment of vibration that is interpreted as sound. This can only be accomplished if the right signal gets to the brain's higher regions. The ear's job is to translate mechanical vibration into electrical impulse with an encoded signal. Similar to a microphone, vibrations can excite the ear. In the microphone, these vibrations are converted into an electrical signal. However, in the ear, they become electrical impulses that are transmitted by nerves and subsequently processed by the brain's central auditory pathways.

The sensory organ, officially referred to as the cochlea, is paired and located on each side of the skull, with the temporal bones acting as a deep cavity for the organ. The cochlea is responsible for transducing vibration, and a portion of the ear is involved in transferring sound to it. Delicate hair cells carry out the transduction and, upon stimulation, produce a nerve impulse. They are submerged in bodily fluid, which gives them nutrition, oxygen, and energy because they are living things. The majority of sound is carried by air vibrations. The two components of the sound conducting mechanism are the middle ear, which serves as an impedance matching device, and the outer part, which detects sound[12]. Figure 1 shows the detailed structure of the ear.

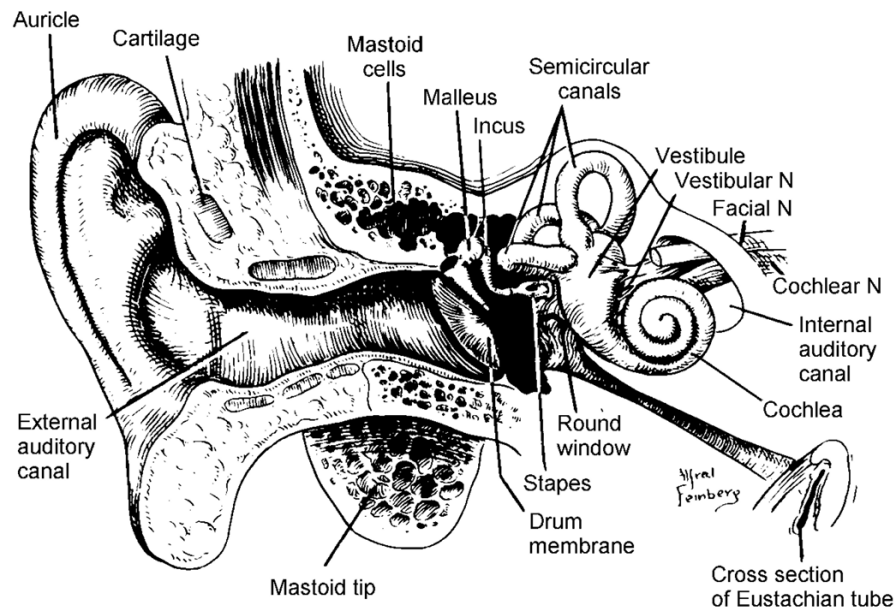


Figure 1: Detailed anatomical structure of the ear.[12]

2.1.1 The outer ear:

The external ear, comprising the pinna and ear canal, plays a pivotal role in sound transmission. The pinna, a cartilaginous structure covered with skin, serves as an acoustic collector, directing sound into the ear canal. Its specific orientation enhances the localization of sounds originating from the front, particularly at higher frequencies. In contrast, the size of the head in relation to sound wavelengths leads to a sound shadow effect in mid frequencies. Lower frequencies are localized through the phase difference in sound arrival between the ears. The ear canal, about 4 centimeters long, consists of an outer segment lined with skin containing hair, sweat glands, and sebaceous glands, responsible for producing earwax. This protective barrier directs sound to the tympanic membrane located at the base of the deeper ear canal, which is firmly attached to the bone and minimizes sound absorption while ensuring efficient sound conduction[12].

2.1.2 The middle ear:

The middle ear, an air-filled cavity, is connected to the nasopharynx by the Eustachian tube, which regulates air pressure. Within the middle ear reside three vital auditory bones: the hammer (malleus), anvil (incus), and stirrup (stapes), responsible for transmitting sound vibrations from the tympanic membrane to the inner ear. The tympanic membrane forms the outer boundary of the middle ear, while the cochlea constitutes the inner boundary. Above, the middle ear is adjacent to the bone beneath the

middle brain lobe, while below, it overlies the jugular bulb, a significant blood-draining vein.

The Eustachian tube's anterior opening and the posterior connection to the mastoid air cells within the temporal bone are located at the front and rear ends of the middle ear, respectively. The middle ear can be envisioned as a pan-like structure with an inclined handle, resembling a frying pan, with the handle indicating the Eustachian tube and an opening in the rear wall leading to the mastoid air cells.[12]

The middle ear extends from the nasal and sinus respiratory spaces, featuring a respiratory membrane that is thick near the Eustachian tube and becomes progressively thinner towards the mastoid region. This space also possesses mucus-secreting capability. The Eustachian tube transitions from bony near the ear to cartilaginous and muscular in the nasopharynx. Muscle contraction actively opens the Eustachian tube to equalize air pressure between the middle ear and the nasal space. Sound transmission occurs through the malleus, which is club-shaped with its handle embedded in the tympanic membrane, connecting it to the incus. The incus, cone-shaped, forms an articulation with the head of the malleus and extends rearward, featuring a thin projection called the long process. The stapes, the third bone, is arch-shaped with a footplate, covering the oval window. This opening leads into the vestibule of the inner ear, known as the cochlea, with the stapediovestibular joint serving as the point of articulation.[12]

2.1.3 The inner ear:

The labyrinth, a bony structure, encompasses the intricately configured membranous canal that constitutes the inner ear. The primary role of the inner ear encompasses two significant functions. The first pertains to the conversion of acoustic pressure into neurochemical impulses, which is executed by the auditory (eighth cranial) nerve and predominantly takes place within the cochlea. The second and third functions are concerned with the preservation of stable visual fixation during motion and the facilitation of an upright posture. These functions are mediated by the vestibular system, a crucial component of the inner ear's complex functionality[13]. Figure 2 shows the general structure of the inner ear for mammals.

The cochlea:

Nestled within the temporal bone, the cochlea is a coiled structure like a snail shell and a vital part of the inner ear. This complex organ converts mechanical vibrations into electrical signals that are perceived by the brain as sound. From a structural perspective, it is made up of three fluid-filled canals that are divided by membranes: the scala vestibuli, scala medium, and scala tympani.[15]

The helicotrema is where the scala vestibuli and scala tympani, which are located at the base and apex of the cochlea, respectively, converge to facilitate communication. The sensitive organ of corti, which includes hair cells essential for hearing perception, is located in the scala media, sometimes

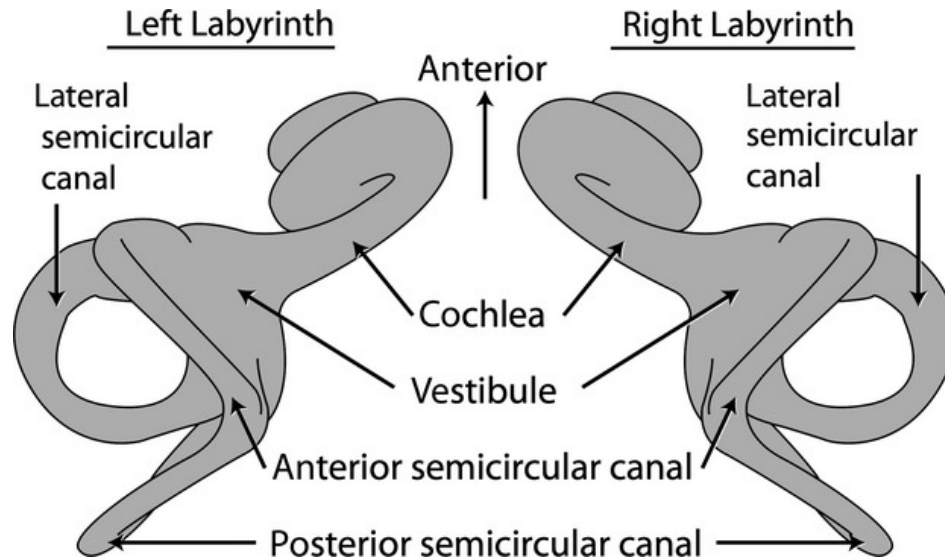


Figure 2: Line drawings of left and right bony labyrinths of a generalized mammal in dorsal view[14].

referred to as the cochlear duct. Endolymph, a specialized fluid rich in potassium ions, envelops this membrane structure lined with sensory cells.

The stapes, which is located near the oval window, causes fluid movements in the cochlea in response to vibrations that travel along the ossicular chain. In order to maintain the equilibrium of the cochlear fluid, the round window at the base of the cochlea functions as a pressure regulator by adjusting for fluid fluctuations brought on by the motions of the oval window.[15]

Important structures surrounding the cochlea include the vestibule, which houses the oval window, the attachment location of the stapes footplate, and the vestibular system, which supports balance and auditory function. Furthermore, the cochlear wall's stria vascularis controls the distinct ionic makeup of endolymph, which is essential for the transmission of auditory signals. The complicated processes of sound perception and cochlear homeostasis maintenance are made possible by the inter-connectivity of these components.

The semicircular canals:

Three semicircular canals exist on either side of the head, aligned in perpendicular planes: the anterior, posterior, and lateral canals. They converge at the utricle, creating five openings, as the anterior and posterior canals unite. At the opposite ends, the anterior and posterior canals form an ampulla, while the lateral canal features an ampulla near the anterior canal's location.[13]

These canals, linked to the utricle, possess sensory hair cells within ampullae that respond to head movements. As the head rotates, the movement of the fluid within the canals triggers hair cell deflection, generating signals carried by the vestibular nerve. These signals aid in maintaining visual fixation

despite head movement through the vestibulo-ocular reflex. Additionally, these canals contribute to the body's balance and posture control, playing a role in the vestibulo-spinal reflex, which aids in postural stability during motion[13]. Figure 3 describes the inner ear anatomical structure as it appears in the axial CT scan.

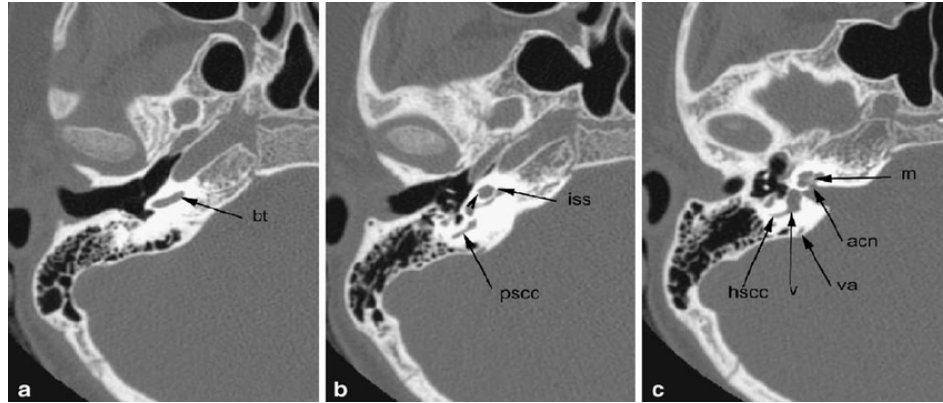


Figure 3: Normal inner ear anatomy demonstrated on axial CT images of the right temporal bone. a Normal basal turn of the cochlea (BT). b A more cephalad image reveals the entirety of the cochlea. The interscalar septum (ISS) between the basal and middle turns is thicker than the septation between the apical and middle turns (arrowhead). The posterior semicircular canal (PSCC) is also seen at this level. c Farther cephalad, the middle and apical turns of the cochlea are shown with the central osseousmodiolus (M). The aperture for the cochlear nerve (ACN) lies between the base of the modiolus and the internal auditory meatus. This image also reveals the vestibule (V), horizontal semicircular canal (HSCC) and vestibular aqueduct (VA).[16]

2.2 Inner ear malformations IEMs:

IEMs are classified into eight distinct groups. most of the distinct characteristics of each group can be seen from the mid-modular trunk view of the inner ear[17].

Normal anatomy NA:

The posterior semicircular canal (PSCC) and a pronounced basal turn are characteristics of the NA inner ear. There is some difference between the middle and apical turns after the basal turn. Because of its reduced breadth, the vestibular aqueduct (VA) is typically not visible in the axial plane. The interscalar septum between the middle and apical twists should be clearly visible. To distinguish between incomplete partition type II and cochlear hypoplasia (CH), this view is crucial.[1]

Cochlear hypoplasia CH:

It is evident how the vestibule and cochlea differ in this deformity. A collection of cochlear malformations known as CH are characterized by a variety of internal architecture deformities and external dimensions

that are smaller than those of a normal cochlea. It is typically challenging to count the turns in a smaller cochlea using CT and/or MRI. However, hypoplasia should be classified as "cochlea with 1.5 turns," not IP-II cochlea. There are now four distinct types of CH defined: CH-I (Bud-like cochlea), CH-II (Cystic hypoplastic cochlea), CH-III (Cochlea with less than 2 turns), CH-IV (Cochlea with hypoplastic middle and apical turns).[1]

Common cavity CC:

A single, ovoid, or round chamber that symbolizes the vestibule and cochlea is known as a CC. This structure contains vestibular and cochlear structures in theory. There could be basic components or accompanying SCC. Usually, the IAC enters the cavity from the center. Sometimes, cases with vestibular dilatation are referred to as "vestibular common cavity," but this is incorrect terminology. This particular malformation type is distinguished by the way the cavity is divided into the vestibular and cochlear portions in the axial view by a line that is drawn parallel to the posterior margin of the internal auditory canal (IAC) in the outer contour.[1, 17]

Incomplete partition type I IPI:

About 20% of IEMs are made up of these. It is evident in this anomaly how the vestibule and cochlea differ from one another. The cochlea appears to be an empty cystic structure because it lacks the entire modiolus and interscalar septa and is situated in the usual location in the anterolateral portion of the IAC fundus. An IP-I cochlea's external measurements (length and height) resemble those of typical cases. Vestibule dilation and enlargement are associated with cochlea. Rarely does vestibular aqueduct enlargement occur. Due to abnormal CA development and the lack of a modiolus, there may be a defect between the IAC and the cochlea, and the cochlea may be fully filled with CSF.[1]

Incomplete partition type II IPII:

The confluence of the middle and apical turns at the apex of the cochlea gives the appearance of a cyst because the apical portion of the modiolus and the corresponding interscalar septa are defective. The cochlea's external measurements, including its height and diameter, are comparable to those of typical cases. Consequently, defining this anomaly as a cochlea with 1.5 turns is incorrect. Only cochleas with 1.5 turns should be referred to as such.[1]

Incomplete partition type III IPIII:

Interscalar septa are present in IP-III cochlea, but the modiolus is completely absent. An anomaly called IP-III cochlear malformation is what causes X-linked deafness. The most uncommon type of

incomplete partition cases is this anomaly. IP-III accounts for two percent of IEMs. Compared to a normal cochlea, the cochlear otic capsule surrounding the membranous labyrinth is thinner in IP-III cochleas. According to the high-resolution CT (HRCT), the otic capsule surrounding the cochlea in IP-III is thin and resembles the membranous labyrinth's outline, seemingly formed by a thick endosteal layer. The second and third layers—instead of the typical three—are most likely either nonexistent or extremely thin. Without enchondral and outer periosteal layers, the innermost endosteal layer seems to be thicker.[1]

Enlarged vestibular aqueduct syndrome EVAS:

This describes the situation where there is a normal cochlea, vestibule, and SCCs but an enlarged vestibular aqueduct (i.e., the midpoint between the posterior labyrinth and operculum is larger than 1.5 mm). The cochlea and vestibule on HRCT and MRI are completely normal, which is a difference between IP-II and EVA. When the midpoint on axial sections between the operculum and posterior labyrinth is greater than 1.5 mm, EVA is traditionally described. The PSSC is clearly visible in the outer contour of EVAS, which was taken from the mid-modular section. The basal turn and some differentiation between the middle and apical turns are also clearly visible. The axial view clearly shows that the VA is wider than the PSSC width.[1, 17]

Ossifications OS:

Cochlear ossification is a pathological process that results in the formation of bony deposits in the cochlea. This condition can be caused by a number of different things, including trauma, infection, and certain autoimmune diseases. In some cases, the exact cause of cochlear ossification is unknown. In axial CT scan it resembles a bony structure overlaying the labyrinth.

Cochlear aplasia CA:

A cochlea-deficient person has cochlear aplasia. The cochlea's typical position is taken up by the labyrinthine portion of the facial nerve, which is anteriorly displaced. This is a distinctive characteristic of the CA malformation type. The vestibule and SCCs are in their normal anatomic placement, at the posterolateral region of IAC. The cavity positions further posteriorly and falls beneath the straight line drawn parallel to the posterior boundary of the IAC.[1, 17]

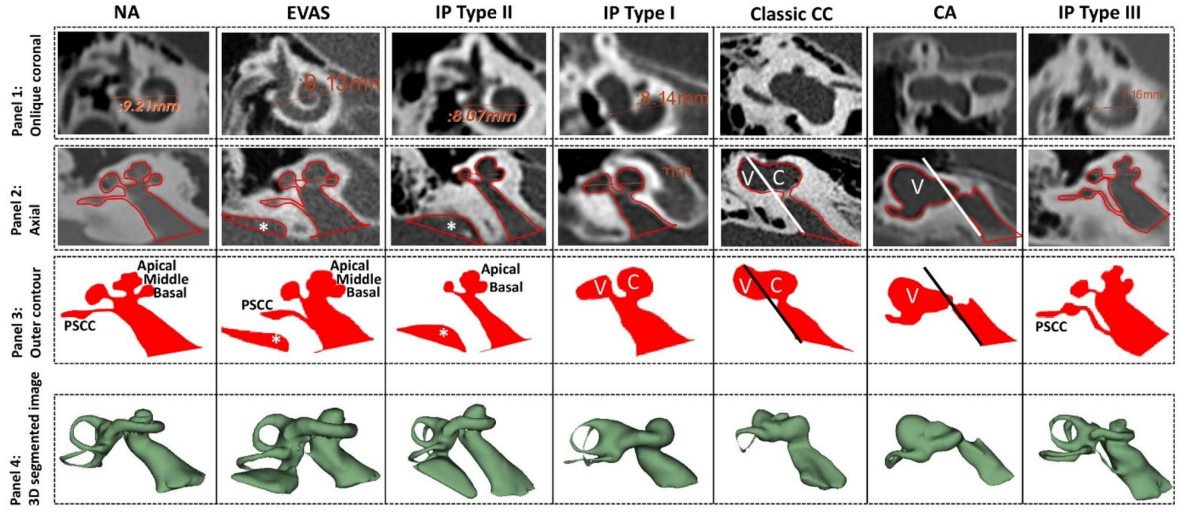


Figure 4: Sample data from every anatomical type in the oblique coronal view is shown in panel 1. Axial view showing the outer contour of the inner ear drawn manually (panel 2) and the captured outer contour filled with color (panel 3). The straight line drawn parallel to the posterior edge of internal auditory canal differentiates the cochlear (C) from the vestibular portion (V) in classic CC and shows only the presence of V in the CA anatomical type. The asterisk (*) refers to the enlarged vestibular aqueduct seen in enlarged vestibular aqueduct syndrome (EVAS) and the incomplete partition (IP) type II. The 3D segmented image of the inner ear allows a better understanding of how the inner ear structures look (panel 4). The 3D segmentation images are not to scale[17].

2.3 K-Means clustering:

One helpful tool in data science is clustering. The greatest similarity within a cluster and the greatest dissimilarity across clusters define the cluster structure that can be found in a data set using this technique. The first clustering technique employed by social scientists and biologists was hierarchical clustering, whereas cluster analysis evolved into a subfield of statistical multivariate analysis.[18]

It is another method of machine learning that uses unsupervised learning. Clustering techniques can be broadly classified into two categories from a statistical perspective: nonparametric approaches and approaches based on probability models. The data points are from a mixture probability model, according to the probability model-based techniques, which means that a mixture likelihood approach to clustering is applied. The expectation and maximization (EM) algorithm is the most popular one in model-based techniques. Clustering methods for nonparametric approaches can be broadly classified as hierarchical or partitional, with partitional methods being the most commonly utilized. These methods are primarily based on an objective function of similarity or dissimilarity measures.[18, 19]

The k-means is one of the most popular unsupervised learning algorithms that solve the well-known clustering problem. Let $X = x_1, \dots, x_n$ be a data set in a d -dimensional Euclidean space R^d . Let $A = a_1, \dots, a_c$ be the c cluster centers. Let $z = [z_{ik}]n \times c$, where z_{ik} is a binary variable (i.e. $z_{ik} \in \{0, 1\}$) indicating if the data point x_i belongs to k -th cluster, $k = 1, \dots, c$. The k-means objective function is $J(z, A) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} |x_i - a_k|^2$. The k-means algorithm is iterated through necessary conditions for minimizing the k-means objective function $J(z, A)$ with updating equations for cluster centers and

memberships, respectively, as

$$a_k = \frac{\sum_{i=1}^n z_{ik} x_i}{\sum_{i=1}^n z_{ik}} \quad (1)$$

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\|x_i - a_k\|$ is the Euclidean distance between the data point x_i and the cluster center a_k . There exists a difficult problem in k-means, i.e., it needs to give a number of clusters a priority. However, the number of clusters is generally unknown in real applications. Another problem is that the k-means algorithm is always affected by initializations.[20].

The percentage of variation explained as a function of the number of clusters is examined using the elbow approach.[21] The foundation of this approach is the notion that one should select a number of clusters such that the addition of another cluster does not result in significantly improved data modeling. A plot is made between the number of clusters and the percentage of variance explained by the clusters. The initial clusters will provide a lot of information, but eventually the marginal gain will sharply decline, creating an angle in the graph. This is the point where the "elbow criterion" is applied, since the right "k" or number of clusters, is selected.[21]

The concept is to calculate your clusters and the training cost by starting with $K = 2$ and increasing it by 1 at a time. The cost decreases considerably at a certain number for K , and when you increase it further, it hits a plateau. This is the desired K value. The reasoning is that after this, you can have more clusters, but some of the current clusters are extremely close to the new cluster. the k-means objective $J()$ decreases quickly as K increases from 1 to 2 and then from 2 to 3. W then reaches an elbow at $K = 3$, after which the distortion decreases extremely slowly. The elbow of this curve indicates that perhaps employing three clusters is the appropriate number of clusters. Because distortion decreases quickly up to $K=3$ and then very slowly thereafter, this data set requires 3 clusters.[21] Figure 5 demonstrates a simple 3D k-means clustering results.[20].

2.4 Convolutional Neural Networks:

Beginning in 2011 and 2012, Convolutional Neural Networks (CNNs)[22] revolutionized the area of neural networks and introduced a new term for it: "deep learning." Many of the most successful deep learning applications, including optical character recognition, speech recognition for smart gadgets,

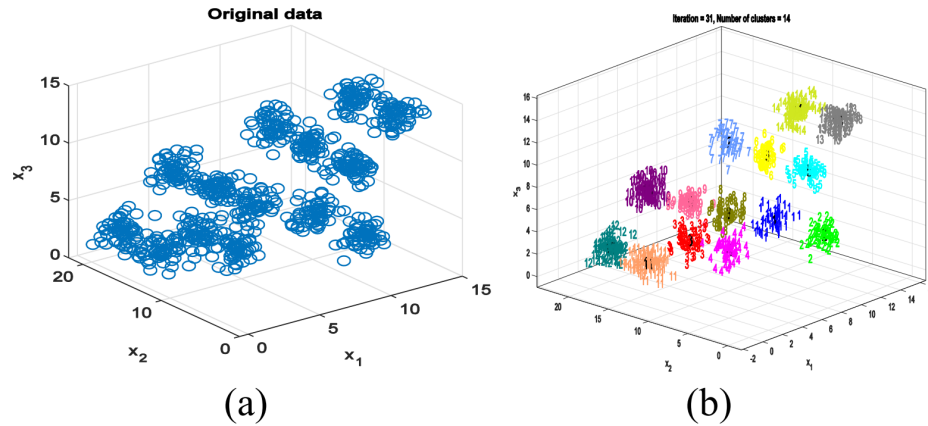


Figure 5: 3D k-means clustering results [20].

and self-driving cars, still rely heavily on CNNs. All of this results from the fact that convolutions are strong yet straightforward tools that assist us in encoding problem information into the architecture of our network. We spend more time engineering our networks' architectures than we do feature engineering.[23]

Convolutions are the usual strategy to utilize for any data that resembles a picture because of their remarkable ability to learn spatial patterns. Applying a convolution to a picture teaches you how to recognize basic patterns such as grid patterns, color shifts, and horizontal or vertical lines. Convolutions are stacked in layers to create increasingly complex patterns by building upon the simpler convolutions that came before them.[23]

Convolution layers are generally regarded as feature extractors. implying that the network should be able to recognize and comprehend more intricate features the more convolution layers the architecture contains. Additionally, the network is able to reduce over-fitting and reach a generalized state in classification with the aid of the dropout layers and batch normalization.[24, 25] Figure 6 demonstrates the general architecture of a 3D CNN.

"Convolutional networks are perhaps the greatest success story of biologically inspired artificial intelligence."

in Goodfellow, Bengio, and Courville 2016[25].

2.4.1 Convolution layers:

The fundamental component of the CNN is the convolution layer. It bears the majority of the computational strain on the network. This layer does a dot product between two matrices: the confined area of the receptive field is one matrix, and the other matrix is the set of learnable parameters, also

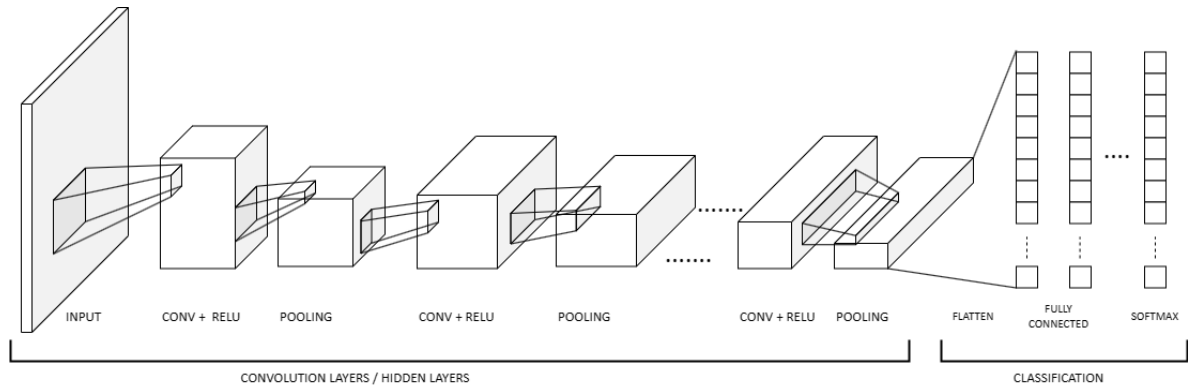


Figure 6: General 3D CNN architecture

referred to as a kernel. Compared to a picture, the kernel is more in-depth but less in space. This indicates that the depth fills up all three channels, yet the kernel height and breadth are spatially little if the image consists of three (RGB) channels.

The kernel creates the image representation of that receptive region by sliding across the image's height and breadth during the forward pass. This results in the creation of an activation map, a two-dimensional representation of the image that shows the kernel's response at each spatial location in the image. A stride is the name given to the kernel's sliding size.

Activation functions:

Non-linearity layers are frequently positioned right after the convolutional layer to add non-linearity to the activation map because convolution is a linear operation, and pictures are not linear.^[25]

- **Sigmoid:** The mathematical version of the sigmoid non-linearity is It converts a real-valued number into a range of values between 0 and 1. Nonetheless, a highly undesirable characteristic of the sigmoid is that the gradient nearly vanishes when the activation is at either tail. In back propagation, an extremely small local gradient will essentially "kill" the gradient. Furthermore, if the neuron receives only positive data, the sigmoid's output will contain either positive or negative values, leading to a spiking dynamic in the gradient updates for weight.
- **Tanh:** Tanh squashes a real-valued number to the range $[-1, 1]$. Like sigmoid, the activation saturates, but — unlike the sigmoid neurons — its output is zero centered.
- **ReLU:** In recent years, the Rectified Linear Unit (ReLU) has gained a lot of popularity. The function is computed. Stated otherwise, the activation is only the threshold at zero. ReLU is more reliable than sigmoid and tanh, and it speeds up convergence by six times. Regretfully, one drawback of ReLU is that it might be brittle during training. It can be updated in such a way

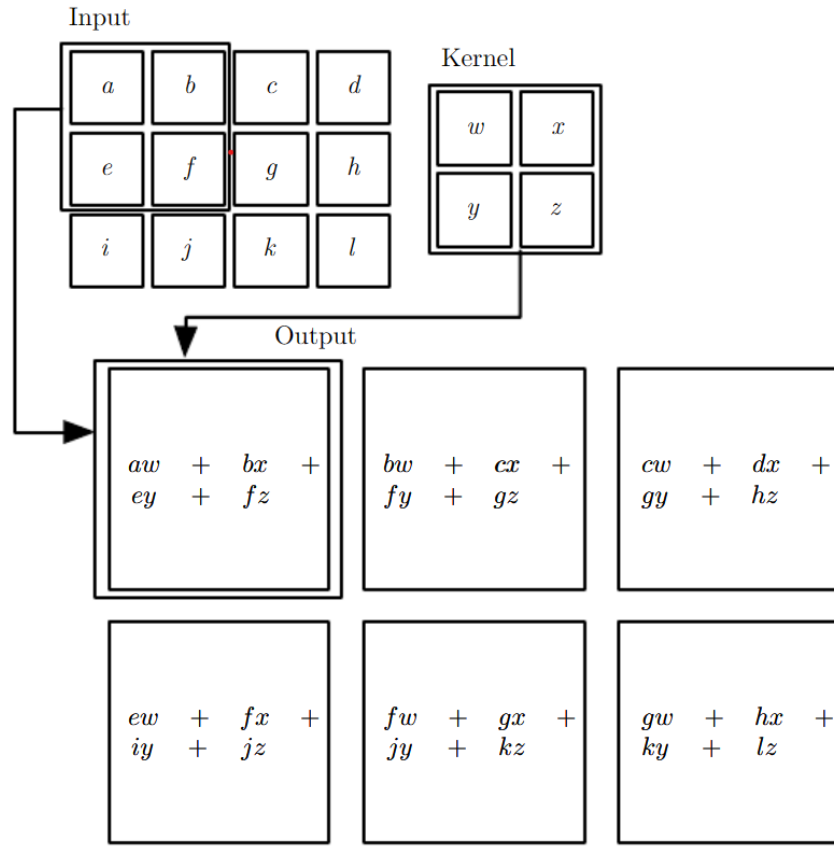


Figure 7: Convolution between a 2x2 kernel and 4x4 input. The weights of the kernel are multiplied with the overlaid pixel values. The result is a feature map representing certain patterns in the input. The resulting feature map is determined by the kernel weights.[25]

that it never receives another update from a huge gradient passing past it. We can deal with this, though, if we establish an appropriate learning rate.

2.4.2 Pooling layers:

By calculating a summary statistic from the outputs in the vicinity, the pooling layer substitutes the network's output at specific points. This aids in shrinking the representation's spatial size, which lowers the quantity of computation and weights needed. Each slice of the representation is processed independently for the pooling operation. The L2 norm of the rectangular neighborhood, the average of the rectangular neighborhood, and a weighted average based on the distance from the central pixel are some examples of pooling functions. Nonetheless, the most widely used method is max pooling, which provides the neighborhood's maximum output.[25]

2.4.3 Fully connected layers:

Neural networks with fully linked layers have an architecture in which every neuron in the layer above it is connected to every other neuron.

Two-dimensional image data and other types of data with a distinct grid-structured topology can be handled by these layers. Their lower computational efficiency in comparison to convolutional networks may have played a role in their previous lack of adoption.

Nonetheless, fully connected networks can handle a wide range of jobs with ease thanks to contemporary hardware. Convolutional networks may have outperformed fully connected ones in the past because of psychological rather than technological obstacles.

It was once thought that fully connected networks did not function well, but with the correct datasets and activation functions, they now seem to function fairly well. Neural networks have gained broader popularity because of the pioneering work of convolutional networks, a specialized type of neural networks. Advances in technology and a change in perspective have made fully connected networks possible again for a variety of uses.[25]

2.4.4 Dropout layers:

Dropout is a regularization approach that, in order to prevent overfitting, randomly "drops out" or "deactivates" a small number of neurons in the neural network.

Instead of training the data on the original network, dropout randomly deactivates the neurons at each training step. The data is then trained on the network with dropped out nodes. The hidden neurons that are silenced by dropout alter in the subsequent training phase iteration due to their probabilistic behavior. Through the use of dropout, or the random deactivation of specific nodes during training, we may effectively replicate an ensemble of neural networks with various designs.[26]

2.4.5 Batch normalization:

The weights assigned to the inputs would have varied greatly if the inputs had been normalized and had distinct ranges for each input since the input features were present in different ranges. Some weights would need to be very large, and some would need to be very little, to account for this range discrepancy across the attributes. The updates related to back-propagation would likewise be huge if we had larger weights, and vice versa. Before the learning algorithm locates the global minima, it continues to oscillate on the plateau region due to this uneven distribution of weights for the inputs.[26]

We standardize the input features such that they are all on the same scale in order to prevent the learning process from spending a lot of time fluctuating on the plateau. The weights assigned to our inputs would likewise be on the same scale because they are on the same scale. Consequently, the

network will train more quickly.

Instead of calculating the mean and standard deviation from the complete set of data, we are doing so from a single batch. Each hidden neuron in the network undergoes batch normalization independently.[26]

Chapter 3

Methodology and Theory:

3.1 Patient cohorts, image analyses, pre-processing:

All experiments were conducted in strict accordance with the Declaration of Helsinki, and the results were reported in full compliance with the Standards for Reporting Diagnostic Accuracy. The retrospective study to use the anonymized images for research was approved by the Institutional Review Board and ethics committees of King Saud University, Saudi Arabia (*IRBNo.22/0084/IRB*), St. Petersburg ENT and Speech Research Institute, Russia (*IRBNo.2301/IRB*), and Stellenbosch University, South Africa (*IRBNo.S2301/IRB*). The ethics committees granted their approval on the grounds that the study was retrospective and did not require informed patient consent.

Three retrospective cohorts from three hospitals/clinics were included in this study. CT scans of the head of 2053 potential cochlear implant (CI) candidates with various inner ear anatomical types were retrieved from the image databases of King Saud University, St. Petersburg ENT and Speech Research Institute, and Stellenbosch University between January 1, 2016, and December 31, 2021. The CT images were acquired using different vendors and scanning parameters, based on the provided Digital Imaging and Communications in Medicine (DICOM) files. Non-enhanced coronal reconstructed images with fields of view that encompassed both ears or one depicted side (Figure 8) were included. In general, 256-slice (range: 128-512) multi-detector-row CT scanners were used. The following scanning parameters were used in the vast majority of cases: axial plane, 0.625 mm (range: 0.1-1.5) slice thickness, auto-mAs (range: 120-145), 140 kV, 484 to 1024 matrix, bone kernel, and rotation time 1 s with 0.3 mm reconstruction in the axial and coronal views. The cases included in the cohorts were selected randomly. Based on the radiological reports, the images were enrolled into six defined categories: (1) normal anatomy (defined as the "normal" group), (2) common cavity, (3) cochlear hypoplasia, (4) enlarged vestibular aqueducts, (5) incomplete partition type I, (6) incomplete partition type III, and (7) ossification (defined as the "abnormal" group). Pre- and post-operative head CTs

of implanted (post-op) and non-implanted ears were also included. The following conditions were excluded: (1) a history of a cranial injury; (2) not sufficient slice thickness (>1 mm) and (3) resolution with more than 0.6 mm/pixel.

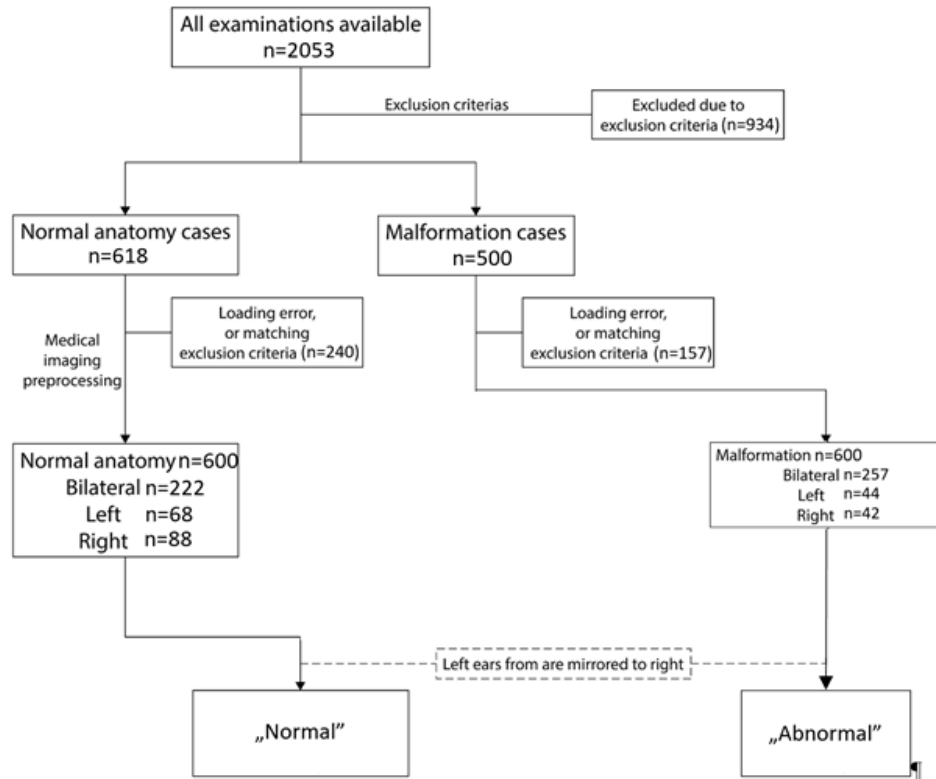


Figure 8: Flowchart outlining the selection of cases. If there were multiple examinations for a patient, the oldest evaluation for abnormalities was included, and the contralateral normal ear was eliminated. Images of the left ear, derived from bilateral images, were all horizontally reversed to the right.

3.2 Cropping algorithms:

We developed different algorithms for cropping the inner ear in a volume and chose the top two for final test:

3.2.1 Hounsfield units and distance-based cropping:

The fundamental principle behind this approach lies in identifying high Hounsfield unit (HU) values within the intricate bone structure of the inner ear. The method initiates by applying an averaging technique using a window filter, typically set at dimensions around 5mm x 5mm x 5mm, in order to merge the inner ear as one blob presented in the whole volume, then it loops through each voxel, starting

from the highest HU value voxel in descending order, then it calculates the distance between each consecutive voxels, if this computed distance exceeds a predetermined threshold, typically set at 20mm, which corresponds to the average longest distance between two points within a normal inner ear, the previous voxel is chosen as the reference point, then a cropping with considering the selected voxel as the center will be applied. Figure 9 describes the algorithm in general.

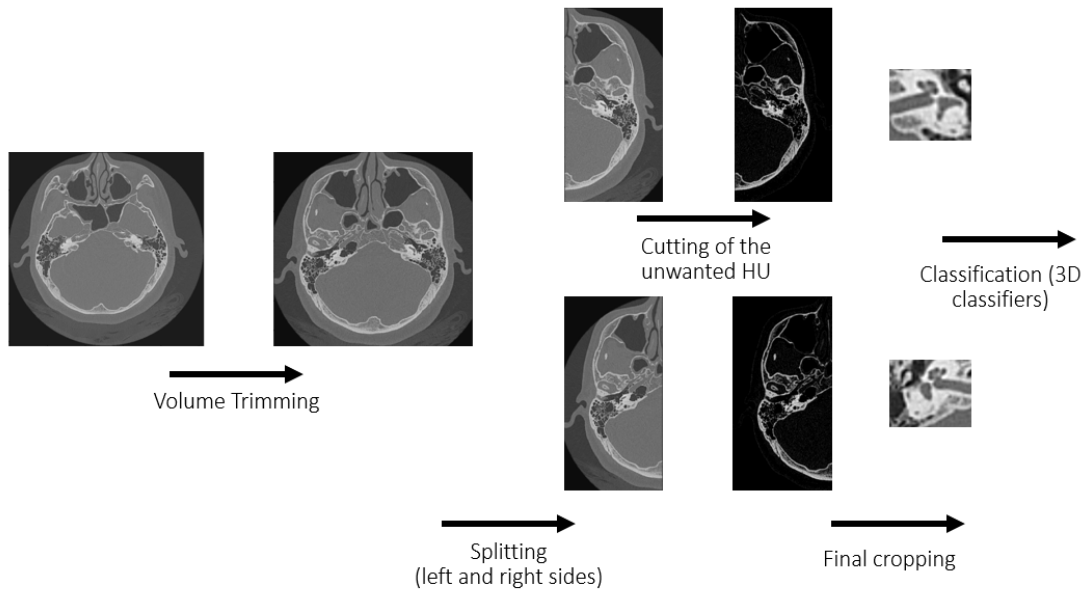


Figure 9: General diagram of Hounsfield units and distance-based cropping algorithm.

3.2.2 Two-staged spatial auto k-means clustering cropping:

This methodology focuses on the interrelations among voxels in addition to their HU values, the first stage starts by averaging the volume with 5mm x 5mm x 5mm kernel, then it takes only the highest 0.5% voxels from the whole volume. The algorithm following this step automatically determines the optimal number of clusters using the elbow method, the next step is to take the closest cluster to the volume center, presuming the inner ear is typically around the volume center, cropping with 45mm x 45mm x 45mm dimensions will be applied, the second stage follows with same procedure but with different parameters namely 3mm x 3mm x 3mm smoothing filter and 20mm x 20mm x 18mm cropping dimensions. The algorithm is summarized in Figure 10.

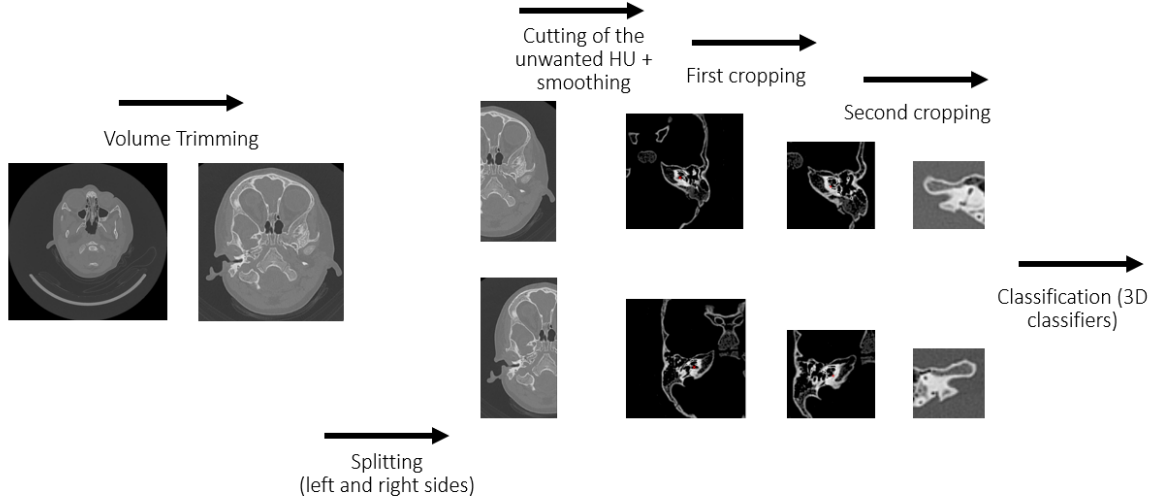


Figure 10: General diagram of Two-staged spatial auto k-means clustering cropping algorithm.

3.3 Dataset, model developments, settings, and evaluation:

Based on the cropped clinical CT series, 600 malformation inputs were selected. 8.33% of the data ($n=50/600$) was saved for secondary validation. Of the remaining 91.67% (550/600), 70% (385/550) was used to train the model and 30% (165/550) was used for testing and validation. The same data pipeline created for normal anatomy ($n=600$). In total, 1200 CT series were selected as input.

Two biomedical engineers with 5 and 13 years of experience, respectively, and a neuro-radiologist with 20 years of experience collaboratively enrolled cases into two labeled datasets: "Normal" and "Malformed." The neural network architecture is shown in figure 11, The evaluation and full data processing pipeline is summarized in Figure 12. We performed a 5-fold cross-validation.

The DL CNN model was created with internal parameters from the ground up. The architecture of this supervised deep learning was as follows: The neural network comprises: (1), Input layer: The input layer has a shape of (width = 128, height = 128, depth = 64, 1), means that the input data is a 3D image with 1 channel. (2), Convolutional layers: The convolutional layers use filters of size (3, 3, 3) to extract spatial features from the input data. The filters are applied to the input data in a sliding window fashion, and the outputs of the filters are combined to produce a feature map. The feature maps are then passed to the next convolutional layer, where more complex features are extracted. (3), Max pooling layers: The max pooling layers downsample the feature maps by taking the maximum value from a pool of pixels. This helps to reduce the computational complexity of the model and also helps to prevent overfitting. (4), Batch normalization layers: The batch normalization layers normalize the outputs of the previous layers, which helps to improve the stability of the training process. (5) Global

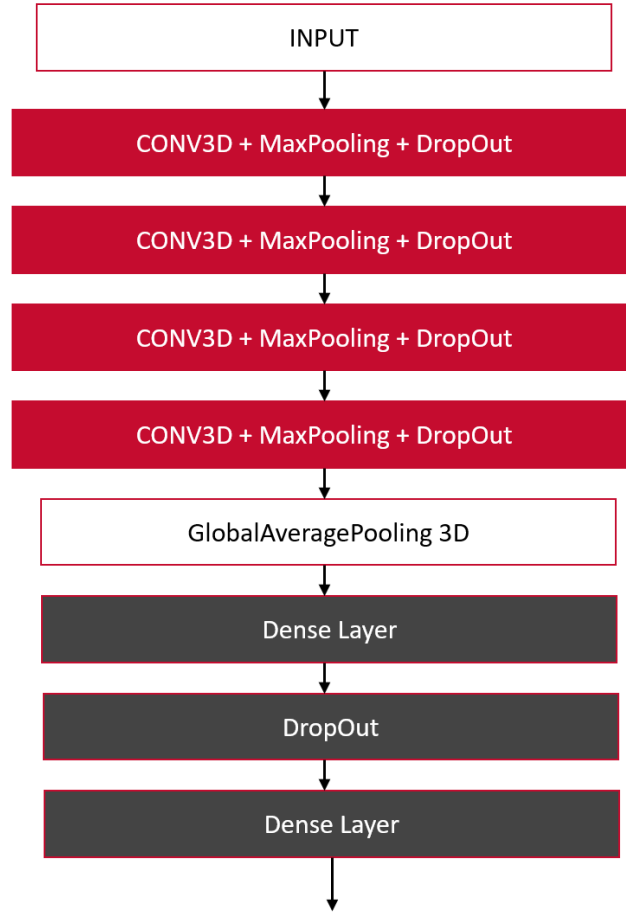


Figure 11: Three-dimensional convolutional neural network model architecture.

average pooling layer: The global average pooling layer averages the outputs of all of the convolutional layers, which produces a single feature vector. (6) Dense layers: The dense layers project the feature vector onto a new space. The first dense layer has 512 units, and the second dense layer has 1 unit. Output layer: The output layer produces a single output value, which is the probability of the input data belonging to a certain class. (7), The parameters of the neural network are optimized and compiled.

After loading the data, it was segregated into defined "Normal" and "Malformed" groups. Then, augmented images for training the CNN were generated. Data augmentation and auto-contrast functions were used each time the training and test data were read, respectively.

The following data augmentation methods were used: random rotation range ($<10^\circ$), random horizontal shifts (-40px - $+40\text{px}$ for three axes), random zoom (65-145%), and random rotation ($<15^\circ$). Additionally, contrast correction was performed by randomly removing the lowest and highest values within 0-2%. Image contrast was altered for the auto-contrast function by removing the 0.7% lowest and

maximum values. All parameters were used to apply random transformations to the images in order to increase the size of the training dataset and improve the robustness of the model. Batch size has been set to 50. For each dataset, the maximum number of epochs was set at 100, and training was terminated early if the validation loss did not decrease after 10 epochs. The models trained on an NVIDIA GeForce RTX 4080 Laptop GPU with 12GB GDDR6 VRAM. We tested the model after training it to see how well it performed.

Gradient-weighted Class Activation Mapping (Grad-CAM) is a popular technique for creating a class-specific heatmap based on a particular input image, a trained CNN, and a chosen class of interest. Grad-CAM is closely related to CAM and it is compatible with any CNN architecture as long as the layers are differentiable. Using this tool, during the prediction, based on the model there are gradients or feature map generated. Feature map represents the found pattern's probability with high intensity, with a color map (rainbow color scale). Where the desired anatomy can be seen, it is highlighted in red, and as we move away from it, the color shifts to blue. Grad-CAM is reimbursed as a heatmap to the original input image. In all cases, Grad-CAM was used to create the interpretation map in order to visualize the basis for the diagnosis.¹² This strategy is ideal for the supervised model because it is independent of CNN structural complexity. We picked up the middle slice (mid-modular section¹³) of selected cases for representative purposes for this research. Qualitative validation based on Grad-CAM.

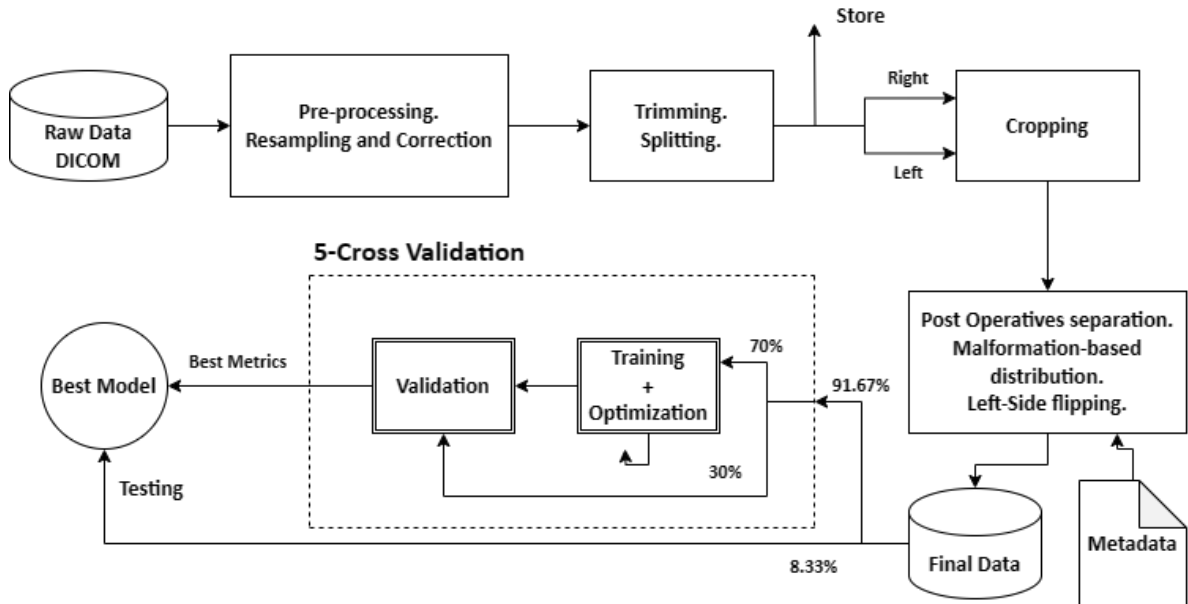


Figure 12: Raw DICOM data is preprocessed and corrected, then it is passed through a trimming and splitting processes to remove the background first and then acquiring the left and right sides accordingly, the resulted data is fed to the cropping algorithm described earlier, cropped images are further corrected and placed in the correct class. All of the shown processes are automatically performed.

3.4 Validation, statistic evaluation:

To validate the developed decision support system, 100 images (50 “Normal” and 50 ”Malformed”) images randomly selected (and separated before model development, and these have never seen by the model during the model training). AI performance for the full dataset (except 100 images for inter-observer analysis) was also measured with standard Area Under the Curve of the Receiver Operating Characteristic curve (AUC-ROC) curve.

All performed statistical analyses, e.g. The Spearman rank-order correlation coefficient, was used to calculate the correlation between two variables. Analyses were carried out using Python 3.12.0.

Chapter 4

Results:

The analysis included 600 “Normal” anatomy and 600 “Malformed” CT series, as can be seen in the flowchart (Figure 12) and evaluated. The most prevalent abnormality was cochlear hypoplasia. The cropping algorithms were evaluated first, and then their speed and usability were improved. Secondly, a command line interface (CLI) program was used to implement them in the data-pipeline. The next step in the process is data preparation. All of the aforementioned functions were then combined under the demo program to illustrate each phase of the process following the training and evaluation of the classification model.

4.1 Cropping algorithms:

Hounsfield units (HU) and distance-based cropping algorithm achieved the following mean accuracy of 87.2% with the malformations. the determined point is more likely to be located around the semi-circular canals.

Two-staged spatial auto k-means clustering cropping algorithm achieved 92.5% mean accuracy with the malformations. The determined point is closer to the basal turn center and Internal auditory canal. Figure 13 shows some samples for the resulting volumes after automatic cropping. Detailed accuracy tables for different malformations are shown in Tables 1, 2 and 3.

Table 1: Accuracy table for the proposed algorithms on the pre operative data

Algorithm	NA	CC	IPI	IPIII	EVAS	OS	CH	Mean Accuracy
HU-based	83%	93%	94%	76%	86%	88%	84%	87.2%
K-means-based	90%	88%	98%	93%	88%	95%	95%	92.5%

Table 2: Accuracy comparison between algorithms on post and pre-operative scans.

Algorithm	Post Operative	Pre Operative
HU-based	41%	87.2%
K-means-based	88%	92.5%

Table 3: Overall cropping accuracies for normal and abnormal anatomies. The abnormal group consists of Common cavity (CC), Incomplete partition type I, Incomplete partition type III (IPIII), Enlarged vestibular aqueduct syndrome, Ossification (OS), Cochlear hypoplasia (CH). Abbreviation: Hounsfield unit (HU)

Algorithm	Normal anatomy	Abnormal anatomy
HU-based	83%	88%
K-means-based	90%	93%

4.2 Data pipeline:

The full data pipeline for a single scan includes: (1) reading scan, (2) resampling to isotropic scan, (3) intensity rescaling, (4) splitting scan to left and right, (5) locating the inner ear in both sides, (6) cropping based on the location, (7) separating post-op and pre-op scans, (8) moving the scan to the correct class based on the metadata.

For processing one scan, the average speed measured was 25 ± 3 sec/scan. The speed depends on the original scan size, original pixel spacing and resolution.

4.3 3D CNN metrics:

During the developments for AI classifier, accuracy, precision, recall, and F1 scores are 0.812, 0.791, 0.8, and 0.766, respectively. Representative scenarios are shown in Figure 14. The supervised DL learning provided accurate categorization in each of the four situations. Performance shown by the area under the curve (AUC) is 0.86 (95% CI: 0.81-0.91). Table 4 summarizes the 3D CNN metrics.

Table 4: 3D Convolutional Neural Network metrics

Metric	Accuracy	Precision	Recall	F1 Score	AUC
Value	0.81	0.79	0.8	0.76	0.86

4.4 OTO-AI demo software:

On top of application programming interfaces (API) and all of the functionalities developed, a graphical user interface (GUI) has been built, it allows the user to load, navigate, and select DICOM scans,

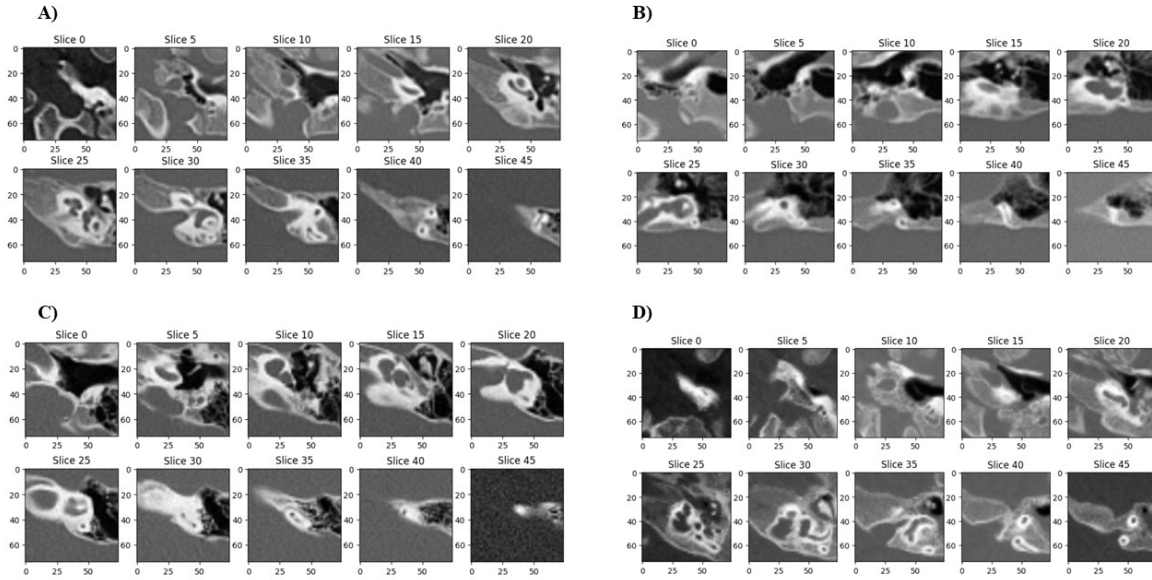


Figure 13: Kmeans cropping algorithm sample results. Most of the resulting volumes contained the whole inner ear structure, although, in some cases, there are dead slices (without the inner ear) and non-centered inner ear volumes (inner ear is not at the center of the volume) A) NA Normal anatomy. B) CC Common cavity. C) IPI Incomplete partition type I. D) IPIII Incomplete partition type III

automatically locating and cropping the ROI using HU-based or K-means-based approach, in addition to manually cropping by locating single marker on the image.

On top of that, a classification can be applied on the cropped side with showing the CI along with classification.

The user can easily navigate between the cropped image and the original one and do the necessary cropping and classification (Figure 15).

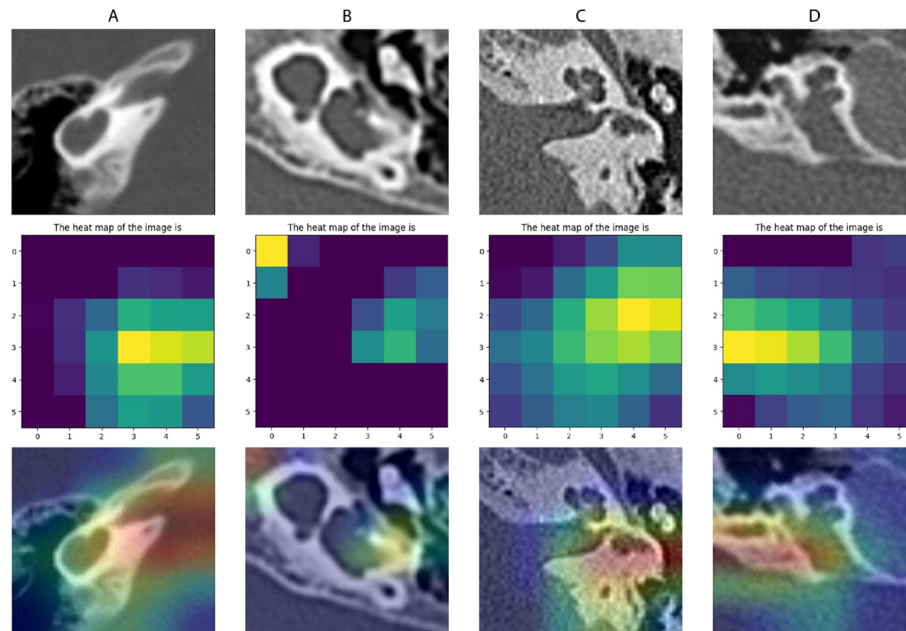


Figure 14: Exemplary instances of the following abnormalities that DL correctly identified: (a) cochlear aplasia with vestibular, (b) Rudimentary otocyst, (c) normal anatomy, and (d) IPIII. On the modular-trunk view CT series the input image, on the second row, the generated jet color GRAD-CAM and the interpolated GRAD-CAM to the CT images represents how the DL decided if the input is IPIII or not. Abbreviations: computed tomography (CT), deep learning (DL), Gradient-weighted Class Activation Mapping (Grad-CAM), incomplete partition type III (IPIII)..

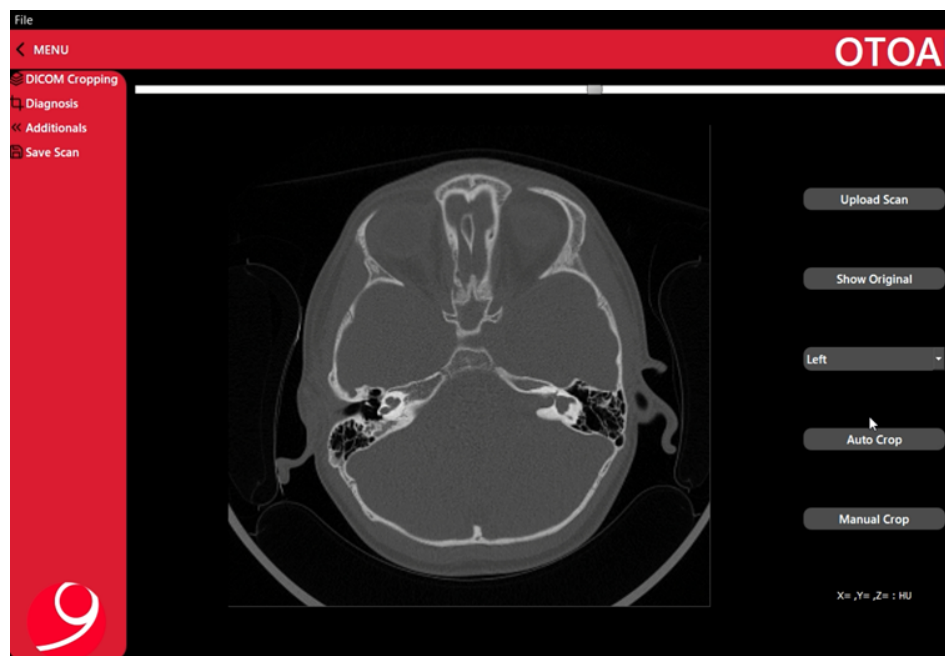


Figure 15: OTOai software prototype GUI along with functions and the classification. Abbreviations: graphical unit interface (GUI) .

Chapter 5

Discussion:

In this study, we developed a method for getting fully automatized and immediate response for inner ear malformation detection, as well as a decision support tool for a complex aspect of inner ear surgery. Two automatized cropping tools developed for cropping the inner ear volume, performances have been measured and compared. Additionally, we created a DL model using a multi-centric, multi-national CT series, and we assessed the models' accuracy, precision, recall, and f1 score as well as their overall performance (AUC). By using cross-sectional medical imaging, our goal was to create an AI-based efficient software with GUI for inner ear diagnosis. Radiology was revolutionized by deep learning, which made it possible to identify anatomical variation.[27] Smaller datasets were used in earlier deep learning-based studies on the diagnosis of malformations in head and neck/temporal bone CT scans.[5, 6, 28]

We assumed that a completely self-developed, easily structured would be suitable for classification, which was also confirmed by the fact in another study with a lot more complex architecture they reached almost the same accuracy.[6] While temporal bone CTs have been the subject of one study, overall, fewer malformations were chosen[5], and the study was limited to a single institute. Similar to that, we developed a supervised model with good learning accuracy[29] however, Ogawa et al. found unsupervised learning to be more promising and achieved the highest accuracy when compared to other supervised learning based studies. This is noteworthy because there is high variance between the two unsupervised learning-based methods for the same purpose, which raises questions.[5, 6]

A previous systematic review by Huang et al. found that deep learning models have achieved diagnostic accuracy comparable to that of human experts in narrow clinical tasks for several medical domains and imaging modalities.[30] Given the complexity of inner ear evaluation, it would be advantageous for end-users to have an assistive AI tool to support their decision-making, especially during the preparation of cochlear implant surgeries. This tool could aid in the selection of the appropriate electrode for a patient-tailored setting and in the prevention of gusher, a complication that

occurs in most incomplete partition type III cases and can lead to inter-operative complications if not anticipated by the ENT surgeon prior to surgery. Fujima et al. also concluded that deep learning techniques may be a beneficial adjunct to the diagnostic process for specific diseases on head CT scans. Fujima et al. also concluded that deep learning techniques may be a valuable counselor to the diagnostic process for specific diseases on head CT scans, and the performance of artificial intelligence is not significantly different from that of human experts.[31] While the goal of this tool is to assist the average user in the absence of a radiological specialist, it would be interesting to conduct future comparisons with the DL model involving a group of radiologists and experts. In our study, we used diversification to reduce risk, based on the Nobel Prize-winning study by Markowitz, who first described that returns can be increased while keeping risk low by diversifying a portfolio.[31] Thus, we covered the variability of input images of different institutes and countries which could be beneficial for young residents and physicians prior to cochlear implant surgery. However, the strict requirements of the General Data Protection Regulation (GDPR) make it difficult for other parties to provide images, so using the tool locally may be a viable approach to obtaining additional expert opinions from outside of the institute. Another possible implementation of our software is for CT vendors, who can offer this initial classification along with the CT machine. If the general head CT examination-based AI indicates a malformation, the CT operator may perform immediately during the same examination a targeted complimentary temporal bone CT scan to better diagnose inner ear disease.

Still, in a world of quickly developing medical imaging technologies, classification systems are limitless. We had 1200 cases from three countries, which could indicate the involvement of more sites and number of images to make a robust model. There are also limitations of the cropping algorithms: HU-algorithm: detecting the teeth when are highly prominent in the scan, and parts of the skull in few cases. K-means algorithm: the detection of the skull in case the inner ear is far from the volume center sometimes happened. Perhaps we could have achieved higher specificity and sensitivity as other studies if we were able to use unsupervised 3D techniques or 3D-Region-Based Convolutional Neural Network (RCNN).[5] AUC values were, however, varied in different research that used the same unsupervised methodology. In the next study, the software should be validated by human experts. Also future research directions shall include the development of multi-class classifiers for additional categories of malformations, auto-segmentation, and/or AI-based methods for cochlear nerve research using magnetic resonance imaging diffusion tensor imaging (MRI DTI). Deep learning methods have the potential to be a valuable tool for the identification of inner ear malformations on head CT scans. This technology could serve as an effective consultant to the diagnostic process.

Chapter 6

Acknowledgments:

We thank all contributors and investigators from MED-EL Medical Electronics GmbH, Department of Research & Development, Innsbruck, Austria. King Saud University, St. Petersburg ENT and Speech Research Institute, Stellenbosch University, University Institute for Diagnostic and Interventional Neuroradiology, University Hospital Bern, University of Pecs, University of Sopron supporting this study. And finally everyone who helped make the research possible.

Chapter 7

Appendix 1:

7.1 Supplementary methods:

7.1.1 Institution overview:

- **Institute 1** Dataset A is collected from an Institute in Russia. This provides diagnostic and rehabilitation of hearing impairments and approximately 400 cochlear implantation are performed at the institute per year.
- **Institute 2** Dataset B is collected from a hospital in Saudi Arabia. This provides tertiary care with 11 departments and 1500 beds.
- **Institute 3** Dataset C is collected from a hospital in South Africa. This provides tertiary case with 30 departments and 1300 beds.

Definition of the training, validation, and test dataset.

The training, validation, and test datasets are defined. The training phase and the test phase are the two main stages of deep learning. For the training stage, it is typical to create a test dataset as well as training and validation datasets 1,2 . To prevent overestimating or overfitting the model, the same data and pictures should not be utilized in any two datasets. Deep learning training often takes place in batches (subsets) that are randomly selected from the training dataset and are then used to optimize the model weights through backpropagation. The model parameters are regularly updated throughout training to get the best possible data fit. The validation dataset is often used to create stopping conditions for the training as well as for parameter selection and tuning. In order to verify the generalizability of the model, it is crucial to assess the classification performance of the artificial intelligence model using an

independent dataset during the testing phase. By dividing the training dataset, the validation dataset is produced using the cross-validation method. In k fold cross-validation, the training dataset is divided into several data subsets k times, with each iteration using a new subset as the held-out validation set. The validation dataset is used to evaluate the performance of the model developed using the training data, whereas the training dataset is used to train the model. The test dataset, albeit not utilized in the training phase, is a separate dataset used to assess model performance and is received from the same facility as the data used in the training phase. Secondary validation is a buffer to measure the performance twice.

7.2 Supplementary references:

1. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019; 380: 1347–58.
2. Liu Y, Chen P-HC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. JAMA 2019; 322: 1806–16.

7.3 Supplementary figures:

7.4 Supplementary codes and usage:

7.4.1 Command Line Interface usage:

usage:

```
IE Processor [-h] -i [str(dir)] [-o [str(dir)]] [-m mal type [mal type ... ]]
              [-s [int(size)]] {full_crop_proc,pre_proc} ...
```

App is intended for processing on CT and MRI. Cropping, Splitting, Resampling.

options:

```
-h, -- help                show this help message and exit
-i [str(dir)], -- input [str(dir)] path to input folder
-o [str(dir)], -- output [str(dir)] path to output folder
-m mal type [mal type ... ], -- malf mal type [mal type ... ]
type of malformation to work with
```

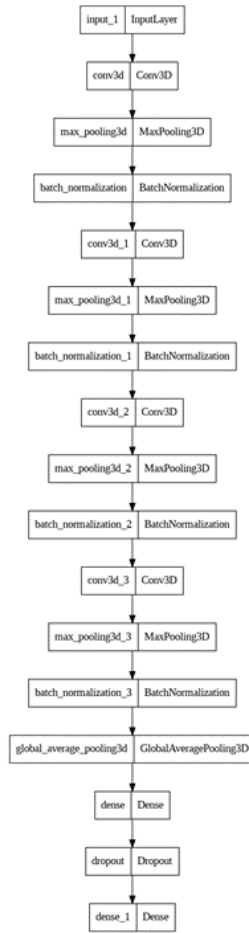


Figure 16: Deep learning convolutional neural network model architecture.

`-s [int(size)], -- size [int(size)]` the sample size per malformation

Processes:

<code>{full_crop_proc,pre_proc}</code>	Different options to manipulate Inner Ear Volumes
<code>full_crop_proc</code>	Apply full process on raw CT to get cropped volumes
<code>pre_proc</code>	Preprocessing of DICOMs

Developed by Abd Alkojak

7.4.2 Inner ear localization and cropping:

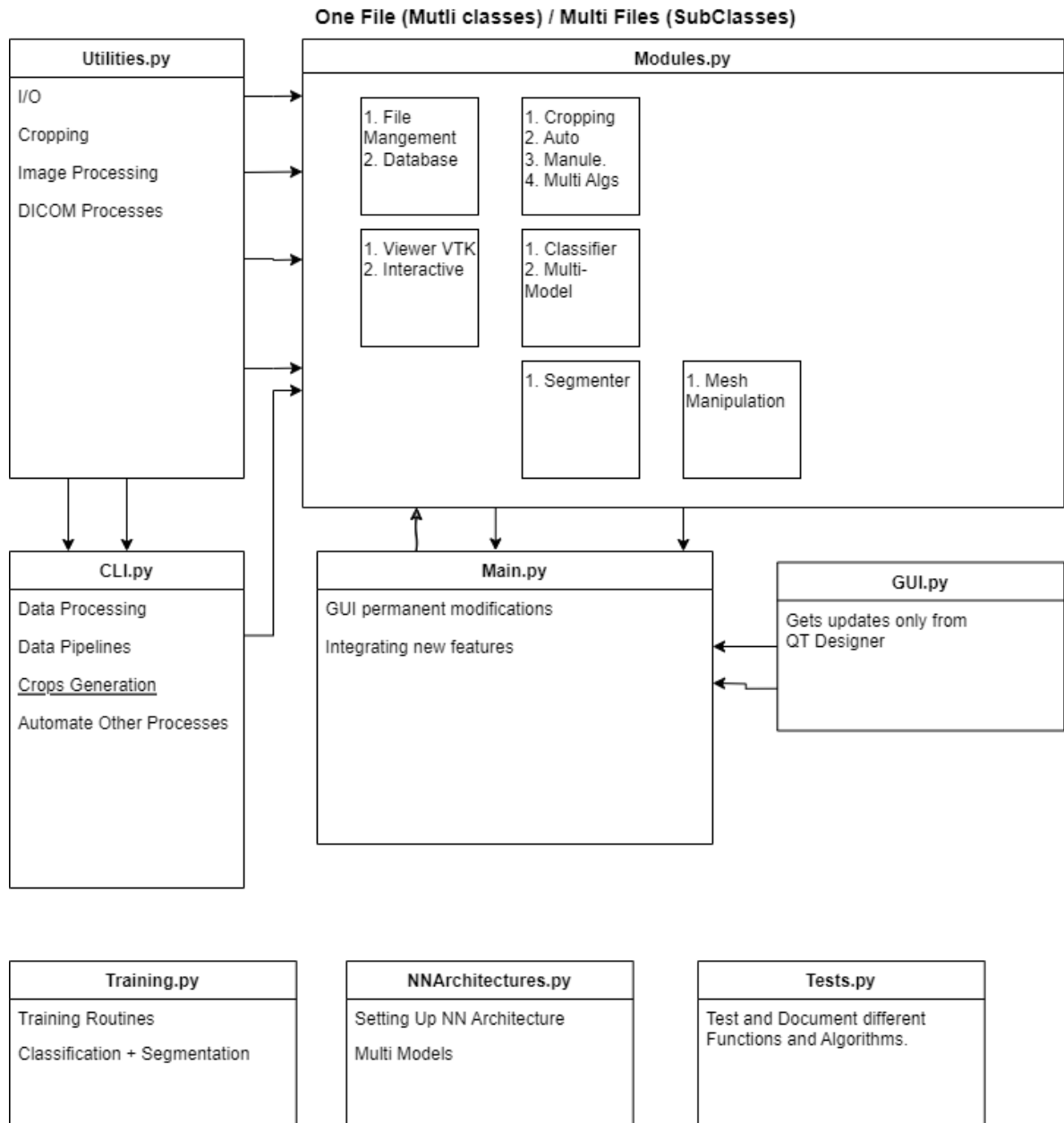


Figure 17: Full project architecture

Algorithm 1 Detect the highest HU blob center with k-means**Require:** image: The DICOM scan to process**Require:** sigma: standard deviation of smoothing filter**Ensure:** final_center

```

1:
2: final_center ← empty list
3: pixel_spacing ← image.Spacing
4: size ← image.Size
5: image_center ← size/2
6:
7: image_hu_cut ← HU_reset(image, range(1000,3071))
8: image_blurred ← SmoothingRecursiveGaussian(image_hu_cut, sigma)
9:
10: max_hu_index ← argsort_reversed(image_blurred_np, all axis)
11: max_hu_index ← max_hu_index 0.05
12: max_hu_coords ← unravel_index(max_hu_index, image_blurred.size)
13: kmeans_data ← array(max_hu_coords)
14: opt_num_clusters ← determine_ncluster(kmeans_data)
15:
16: kmeans ← Kmeans(n_clusters=opt_num_clusters, n_init= auto)
17: kmeans.fit(kmeans_data)
18: cluster_center_closest_index ← empty list
19:
20: for cluster in range (opt_num_clusters) do
21:     cluster_center ← kmeans.cluster_centers_[cluster]
22:     cluster_center_closest_index ← append(argmin(norm(cluster_center)))
23: end for
24:
25: cluter_center_coords ← kmeans_data[cluster_center_closest_index]
26: cluter_center_coords ← empty list
27: for point in cluter_center_coords do
28:     if !point_inside_margins(image_blurred, point) then
29:         cluter_center_coords ← append(point)
30:     end if
31: end for
32:
33: if len(cluter_center_coords) < 1 then
34:     final_center ← center
35: else
36:     final_center ← filter_regions_area_maxis(image, cluter_center_coords,
        sigma)
37: end if
        return (final_center)

```

Algorithm 2 Helper function: Filter regions based on area and minimum axis

Require: image**Require:** center_list: list of x,y,z coordinates**Require:** pixel_spacing**Require:** sigma: standard deviation of smoothing filter**Ensure:** final_center

```
2: minor_axis_list ← empty list
   image_blurred ← SmoothingRecursiveGaussian(image,sigma)
4:
   for center in center_list do
6:   image_cropped ← crop_from_center(image_blurred, center, 22x22x15 mm)
   image_cropped ← OtsuThresholdImageFilter(sitk_cropped)
8:   st_size ← 0.5 mm / pixel_spacing
   footprint ← ball(st_size)
10:  image_euler ← opening(image_cropped,footprint)
   labeled_mask ← getlabel(image_euler)
12:  labeled_mask ← remove_small_objects(labeled_mask,min_size=10)
   props ← regionprops(labeled_mask)
14:  largest_region ← max(props, key= region.area)
   minor_axis_list ← append(largest_region.minor_axis_length)
16: end for

18: final_center ← center_list[argmax(minor_axis_list)]
   return final_center
```

Algorithm 3 Helper function: Check if point is inside image margins

Require: image: image as nd array**Require:** point: x,y,z coordinate**Require:** margins: edges of the volume calculated as coordinates**Ensure:** True if point is inside margins, False otherwise

```

    vol_margins  $\leftarrow$  empty dictionary
3:   for i, (dim , perc) in enumerate(zip(["z","y","x"],margins)) do
        vol_margins[dim]  $\leftarrow$  [perc image.shape[i], image.shape[i] - perc image.shape[i]]
6:   end for

    for p, m in (point, vol_margins.values()) do
9:       in_range  $\leftarrow$  append (p <= m[0] or p >= m[1])
    end for

12: if any of the content is false inside in_range then
        return True
    else
        return False
    end if

```

Algorithm 4 Helper function: Determine the number of clusters using elbow method and rate of change

Require: data: data point representing voxel values**Require:** max_clust_num: maximum number of clusters**Require:** extra_cluster: flag to increase cluster numbers by 1**Ensure:** possible_clusters[elbow_index]

```

    inertia_values  $\leftarrow$  empty list
    possible_clusters  $\leftarrow$  range(1,max_clust_num+1)
4:   for n in possible_clusters do
        kmeans  $\leftarrow$  initialize Kmeans with n clusters
        kmeans  $\leftarrow$  training(data)
8:   inertia_values.append(kmeans.inertias)
    end for

    rate_of_change  $\leftarrow$   $\Delta$ inertia_values /  $\Delta$ possible_clusters
12: elbow_index  $\leftarrow$  index where(rate_of_change < 0.01) + Toint(extra_cluster)
    return possible_clusters[elbow_index]

```

Algorithm 5 Full Processing Pipeline

Require: image: the input medical image

Require: spacing: the desired spacing for resampling

Require: crop_volume: the volume dimensions for cropping

Ensure: (image_left_cropped, image_right_cropped): the processed and cropped left and right inner ear sides

Step 1: Resample to Isotropic with Standard Axes

image_resampled $\leftarrow$ resample_to_isotropic(image, spacing, standard)

5: **Step 2: Reset the Hounsfield Units (HU)**

image_reset $\leftarrow$ HU_reset(image_resampled, range=(-1024, 3071))

Step 3: Trim the Image Based on the Skull

(image_trimmed, flag) $\leftarrow$ trim_volume_HU_based(image_reset)

10:

Step 4: Split the Image into Left and Right Halves

(image_left, image_right) $\leftarrow$ split_to_left_and_right(image_trimmed)

Step 5: Crop the Left and Right Images to Square Shapes

15: image_left $\leftarrow$ crop_to_square(image_left, size(image_left))

image_right $\leftarrow$ crop_to_square(image_right, size(image_right))

Step 6: Detect the Highest HU Blob Center Using K-means

center_left $\leftarrow$ detect_highest_HU_blob_center_kmeans(image_left)

20: center_right $\leftarrow$ detect_highest_HU_blob_center_kmeans(image_right)

Step 7: Crop the Images Around the Detected Centers

image_left_cropped $\leftarrow$ crop_from_center(image_left, crop_volume, center=center_left)

image_right_cropped $\leftarrow$ crop_from_center(image_right, crop_volume, center=center_right)

25: **return** (image_left_cropped, image_right_cropped)

Bibliography

- [1] L. Sennaroğlu and M. D. Bajin, “Classification and current management of inner ear malformations,” *Balkan Medical Journal*, vol. 34, pp. 397–411, 9 2017.
- [2] B. Sun, P. Dai, and C. Zhou, “[study on 2,747 cases of inner ear malformation for its classification in patient with sensorineural hearing loss].” *Lin chuang er bi yan hou tou jing wai ke za zhi = Journal of clinical otorhinolaryngology, head, and neck surgery*, vol. 29, pp. 45–7, 1 2015.
- [3] G. Widmann, D. Dejaco, A. Luger, and J. Schmutzhard, “Pre- and post-operative imaging of cochlear implants: a pictorial review.” *Insights into imaging*, vol. 11, p. 93, 8 2020.
- [4] M. Bassiouni, H.-C. Bauknecht, G. Muench, H. Olze, and J. Pohlen, “Missed radiological diagnosis of otosclerosis in high-resolution computed tomography of the temporal bone-retrospective analysis of imaging, radiological reports, and request forms.” *Journal of clinical medicine*, vol. 12, 1 2023.
- [5] M. Ogawa, M. Kishihara, T. Yamamoto, S. Shibata, Y. Ojio, K. Mochizuki, A. Tatsuta, S. Iwasaki, and Y. Shibamoto, “Utility of unsupervised deep learning using a 3d variational autoencoder in detecting inner ear abnormalities on ct images.” *Computers in biology and medicine*, vol. 147, p. 105683, 8 2022.
- [6] P. L. Diez, J. Margeta, K. Diab, F. Patou, and R. R. Paulsen, *Unsupervised Classification of Congenital Inner Ear Malformations Using DeepDiffusion for Latent Space Representation*. publisher name, 2023, pp. 652–662.
- [7] A.-T. Radutoiu, F. Patou, J. Margeta, R. R. Paulsen, and P. L. Diez, *Accurate Localization of Inner Ear Regions of Interests Using Deep Reinforcement Learning*. publisher name, 2022, pp. 416–424.
- [8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 12 2017.

- [9] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, “A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis,” *The Lancet Digital Health*, vol. 1, pp. e271–e297, 10 2019.
- [10] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, “3d deep learning on medical images: A review.” *Sensors (Basel, Switzerland)*, vol. 20, 9 2020.
- [11] L. Ruff, R. A. Vandermeulen, B. J. Franks, K.-R. Müller, and M. Kloft, “Rethinking assumptions in deep anomaly detection,” *journal name*, 5 2020.
- [12] P. W. Alberti, “The anatomy and physiology of the ear and hearing,” *Occupational exposure to noise: Evaluation, prevention, and control*, pp. 53–62, 2001.
- [13] J. Irwin, “Basic anatomy and physiology of the ear,” *Infection and hearing impairment*, pp. 8–13, 2006.
- [14] E. Ekdale, “Form and function of the mammalian inner ear,” *Journal of Anatomy*, vol. 228, 04 2015.
- [15] H. Rask-Andersen, W. Liu, E. Erixon, A. Kinnefors, K. Pfaller, A. Schrott-Fischer, and R. Glueckert, “Human cochlea: Anatomical characteristics and their relevance for cochlear implantation,” *The Anatomical Record*, vol. 295, no. 11, pp. 1791–1811, 2012. [Online]. Available: <https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/ar.22599>
- [16] C. Robson, “Congenital hearing impairment,” *Pediatric radiology*, vol. 36, pp. 309–24, 05 2006.
- [17] A. Dhanasingh, D. Erpenbeck, M. Z. Assadi, Úna Doyle, P. Roland, A. Hagr, V. V. Rompaey, and P. V. de Heyning, “A novel method of identifying inner ear malformation types by pattern recognition in the mid modiolar section,” *Scientific Reports*, vol. 11, p. 20868, 10 2021.
- [18] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Technometrics, 1988. [Online]. Available: <https://api.semanticscholar.org/CorpusID:29535089>
- [19] A. K. Jain, “Data clustering: 50 years beyond k-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010, award winning papers from the 19th International Conference on Pattern Recognition (ICPR). [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865509002323>

- [20] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020.
- [21] P. Bholowalia and A. Kumar, "Ebk-means: A clustering technique based on elbow method and k-means in wsn," *International Journal of Computer Applications*, vol. 105, pp. 17–24, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:1759661>
- [22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [23] E. Raff, *Inside deep learning: Math, algorithms, models*. Simon and Schuster, 2022.
- [24] C. Ranjan, "Understanding deep learning: Application in rare event prediction," *Imbalanced Learning; Connaissance Publishing: Atlanta, GA, USA*, pp. 19–27, 2020.
- [25] J. Heaton, "Ian goodfellow, yoshua bengio, and aaron courville: Deep learning," *Genetic Programming and Evolvable Machines*, vol. 19, pp. 305–307, 2018.
- [26] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: an empirical study of their impact to deep learning," *Multimedia Tools and Applications*, vol. 79, pp. 12 777–12 815, 5 2020.
- [27] Z. Hasan, S. Key, A.-R. Habib, E. Wong, L. Aweidah, A. Kumar, R. Sacks, and N. Singh, "Convolutional neural networks in ent radiology: Systematic review of the literature," *Annals of Otolaryngology, Rhinology and Laryngology*, vol. 132, pp. 417–430, 4 2023.
- [28] D. Zhang, J. H. Noble, and B. M. Dawant, "Automatic detection of the inner ears in head ct images using deep convolutional neural networks." *Proceedings of SPIE—the International Society for Optical Engineering*, vol. 10574, 2 2018.
- [29] A. A. H. de Hond, E. W. Steyerberg, and B. van Calster, "Interpreting area under the receiver operating characteristic curve." *The Lancet. Digital health*, vol. 4, pp. e853–e855, 12 2022.
- [30] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines." *NPJ digital medicine*, vol. 6, p. 74, 4 2023.
- [31] N. Fujima, V. C. Andreu-Arasa, K. Onoue, P. C. Weber, R. D. Hubbell, B. N. Setty, and O. Sakai, "Utility of deep learning for the diagnosis of otosclerosis on temporal bone ct." *European radiology*, vol. 31, pp. 5206–5211, 7 2021.