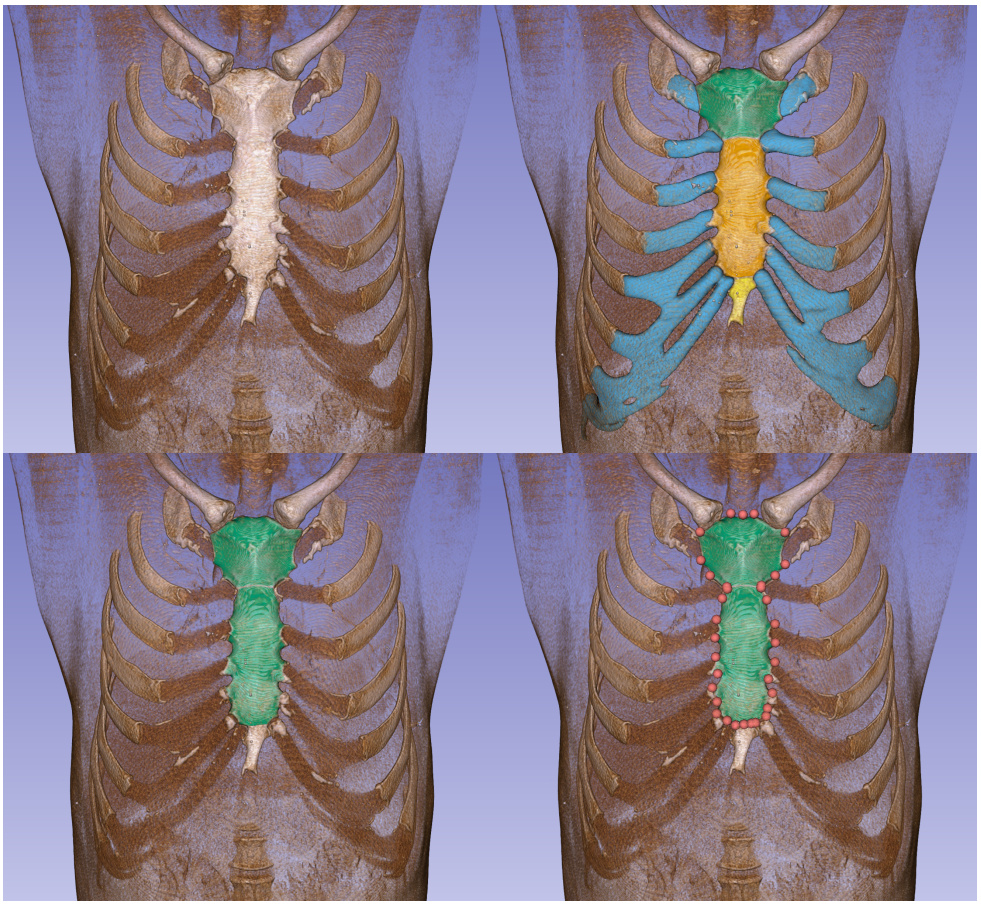


Deep Learning for Segmentation and Landmarking of the Human Anatomy

Applying nnU-Net and SpatialConfiguration-Net to Automate Statistical Shape Modeling for Sternum and Costal Cartilage

Master's thesis in *Complex Adaptive Systems* and *Systems, Control and Mechatronics*

Anton Alexandersson
Anna Molnö

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's Thesis 2025

Deep Learning for Segmentation and Landmarking of the Human Anatomy

Applying nnU-Net and SpatialConfiguration-Net to Automate Statistical Shape Modeling for Sternum and Costal Cartilage

Anton Alexandersson
Anna Molnö

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology
Gothenburg, Sweden 2025

Deep Learning for Segmentation and Landmarking of the Human Anatomy
Applying nnU-Net and SpatialConfiguration-Net to Automate Statistical Shape Modeling for Sternum and Costal Cartilage
Anton Alexandersson
Anna Molnö

Supervisors: Oscar Hallberg and Johan Iraeus, Department of Mechanics and Maritime Sciences
Examiner: Johan Davidsson, Department of Mechanics and Maritime Sciences

Master's Thesis 2025
Department of Mechanics and Maritime Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Cover: Visualization of thoracic anatomy based on a computed tomography (CT) scan, rendered in 3D Slicer. The image series shows (from left to right, top to bottom): the unprocessed thorax, segmentation of the sternum and costal cartilages, isolated sternum segmentation, and the segmented sternum with anatomical landmarks.

Typeset in LaTeX
Gothenburg, Sweden 2025

This report presents the outcome of our master's thesis project carried out at the Department of Mechanics and Maritime Sciences at Chalmers University of Technology during the spring of 2025.

Deep Learning for Segmentation and Landmarking of the Human Anatomy
Applying nnU-Net and SpatialConfiguration-Net to Automate Statistical Shape Modeling for Sternum and Costal Cartilage
ANTON ALEXANDERSSON
ANNA MOLNÖ
Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology

Abstract

During vehicle development, simulations are used in part to replicate the mechanical response of the human body during crash testing. To ensure realism, human body models are employed. However, current models often lack an accurate representation of thoracic anatomy, including the sternum and costal cartilage. One approach to improve these models is to develop a statistical shape model based on computed tomography (CT) scans of the thoracic region.

The creation of statistical shape models from CT data involves multiple steps and is typically time-consuming. This study focuses on automating the first two steps of the process: segmentation and landmarking, using two different deep learning models. The aim is to enhance efficiency and consistency compared to manual processing. Additionally, the impact of dataset size on the models' performance is investigated.

The creation of statistical shape models from CT data involves multiple steps and is typically time-consuming. This study focuses on automating the first two steps of the process: segmentation and landmarking using deep learning. The aim is to enhance efficiency and consistency compared to manual processing. Additionally, the impact of dataset size on model performance is investigated.

The methodology includes preparing datasets with manual segmentations and landmarks, formatting them for use with deep learning models, and training two architectures: nnU-Net for segmentation and SpatialConfiguration-Net (SCN) for landmarking. The segmentation results from nnU-Net are generally accurate and more consistent than manual segmentations, though some manual post-processing is still necessary. SCN also demonstrates promising performance for landmarking, with similar requirements for manual correction. Both models demonstrate improved performance as the size of the training dataset increases. Additionally, nnU-Net's segmentation accuracy improves when trained on datasets with more precise annotations. Despite the small dataset sizes (up to 10 scans for segmentation and 20 for landmarking), both models achieved promising results, highlighting their robustness even with limited data.

In conclusion, deep learning can effectively automate the segmentation and landmarking processes required for statistical shape model development, offering improvements in both efficiency and consistency. The findings also suggest that model performance continues to improve with larger and more diverse datasets.

Keywords: Deep Learning, Automatic Segmentation, Automatic Landmarking, nnU-Net, SpatialConfiguration-Net, Medical Imaging, Thorax, Costal Cartilage, Sternum

Acknowledgements

First of all, we would like to sincerely thank our supervisors, Oscar Hallberg and Johan Iraeus, at the Department of Mechanics and Maritime Sciences at Chalmers, for their support, engagement, expertise, and patience, which made this master's thesis possible. With their guidance and academic knowledge, the project progressed smoothly. We are truly grateful - this thesis would not have been possible without your invaluable support.

We would also like to express our sincere gratitude to Jennifer Alvén and David Hagerman Olzon at the Department of Electrical Engineering at Chalmers for their time and guidance in applying deep learning methods. Jennifer guided us in the work with automatic landmarking, and David generously took time to help us with automatic segmentation. Without their advice and expertise, this project would not have advanced as far or achieved the results it did. Thank you!

Finally, we want to thank all our friends who made our years at Chalmers truly memorable - it has been a blast!

Anton Alexandersson, Anna Molnö, Gothenburg, June 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

CNNs	Convolutional Neural Networks
CSV	Comma Separated Values
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
DNNs	Deep Neural Networks
FE	Finite Element
HBMs	Human Body Models
HU	Hounsfield Unit
IPE	Inter-Point Error
LPS	Left, Posterior, Superior
NMDID	New Mexico Decedent Image Database
NRRD	Nearly Raw Raster Data
RAS	Right, Anterior, Superior
SCN	SpatialConfiguration-Net
SD	Standard Deviation

Nomenclature

Below is the nomenclature of indices and parameters that have been used throughout this thesis.

Indices

i	Index for landmarks

Parameters

d	Dimensionality
γ	Heatmap scaling factor
σ_i	Standard deviation for heatmap i
α	Penalty coefficient for standard deviations
λ	Penalty coefficient for network weights
β_1	Decay rate for the momentum in Adam optimizer
β_2	Decay rate for the squared gradients in Adam optimizer

Contents

List of Figures

List of Tables

1

Introduction

The pursuit of a safe road experience is of interest to the Swedish state as well as to the vehicle companies. Since the late 1990s, the state has strived towards *vision zero*, which is a long-term project aiming for zero cases of severe and fatal injuries. The vision states, among other aims, that the system developers should design vehicles that prevent fatal and severe injuries in traffic [1]. Apart from *vision zero*, vehicle companies also prioritize safety features, since it is one of the most important aspects of vehicle purchasing for consumers [2].

To ensure vehicle safety, both physical and virtual tests are conducted to gain insights into how vehicle accidents impact the human body. Autoliv, Volvo, and Sahlgrenska, along with Chalmers, are engaged in an ongoing research project (FFI 2023-02613) aimed at improving the accuracy and reliability of virtual testing by creating a model of the human body that more realistically represents human anatomy and physiology than current models. More specifically, it seeks to enhance torso injury risk prediction. The goal of this thesis is to support the overarching research project by evaluating two deep learning models with proven performance on small medical datasets, one for automatic segmentation and one for landmarking.

1.1 Background

Numerous physical tests are conducted during vehicle development, including those for safety functions. These crash tests are performed to test the performance of a vehicle in an accident. These tests, while essential, are costly and require physical prototypes. To reduce the number of physical tests, simulations are used [3]. To make the simulations possible, data on the road conditions as well as the driver and occupants are necessary. With data on the driver and occupants, human body models (HBMs) can be created, which are used to simulate the mechanical response of the body during a crash.

The current HBMs in use cannot predict fracture risks in the sternum and costal cartilage, and one of the research aims of FFI 2023-02613 is to enable this. The type of HBMs used in FFI 2023-02613 are finite element (FE) models [4]. The FE models are based on statistical shape models, which in turn are based on geometries from computed tomography (CT) scans [5]. The first step to achieving information about the geometries is to segment the desired anatomical objects in the CT images: the sternum and costal cartilage. To segment a medical image means that specific

regions, such as organs, bones, or tumors, are marked and classified. Once the desired objects are segmented, landmarks can be placed on the object's surface. Landmarks are points that represent the object's geometry. With the landmarks as a frame, a template mesh is shaped to the object. The reshaped mesh holds information about how the object differs from the template object. The information that the reshaped mesh holds is then used as the basis for a statistical analysis, which in turn is part of the statistical shape model [6].

Human anatomy exhibits high variability, and representing this in a statistical shape model requires a large dataset. However, manually segmenting and placing landmarks on the sternum and costal cartilages is highly time-consuming and prone to human error. Therefore, an automated approach is preferable for these steps. One promising method for automatic segmentation and landmarking of 3D images is the use of deep learning models.

1.2 Purpose

The primary purpose of this study is to explore the feasibility and development of a method for automatic segmentation of CT images, intending to achieve a more efficient alternative to manual segmentation. In addition, the study also aims to explore the potential of integrating automatic landmarking as a complementary step. The following research questions guide this work:

- How can a deep learning model be adapted for sternum and costal cartilage segmentation?
- What level of accuracy can be achieved with a limited amount of manually annotated data?
- How does the amount of manually annotated data affect the resulting segmentation?
- How can the approaches and insights from segmentation be extended to automatic landmarking of the sternum?

1.3 Goals

The primary goal of this thesis is to develop a method for automated segmentation of the sternum and costal cartilages in CT images using a deep learning model called nnU-Net. A secondary goal is to explore the concept of automatic landmarking using a deep learning model called SpatialConfiguration-Net. By automating these processes, segmentation and landmarking can be performed more efficiently, allowing for the inclusion of more subjects in the research project FFI 2023-02613. This broader dataset will help capture greater variability within the population, contributing to a more accurate statistical shape model.

1.4 Limitations and demarcations

The segmentation in this study is demarcated to the costal cartilages and sternum in standardized CT images from the New Mexico Decedent Image Database (NMDID) [7]. Landmarking is restricted to the sternum only. A key limitation for both segmentation and landmarking is the small number of manually segmented and annotated samples, which may affect the models' abilities to generalize across a diverse population.

1.5 Ethics

The NMDID consists of decedents that have been CT scanned. There is no consent, informed or implied, associated with the database, as deceased persons are not considered human subjects under US federal laws regarding research on human subjects. Despite this, numerous privacy protections are built-in, and the subjects are considered de-identified [7]. Additionally, this project only used the thorax, which further anonymized the decedents. During the project, all the data used was stored on an external SSD drive. The SSD drive was under supervision during work hours and in safekeeping at all other times so that only authorized persons had access to it.

2

Theory

This chapter presents the theoretical background necessary to understand the methods used in developing automatic approaches for segmentation and landmarking. It introduces key concepts in deep learning, with a closer look at the models used in this thesis: *nnU-Net* and *SpatialConfiguration-Net (SCN)*. In addition, the chapter provides an overview of relevant medical data formats and the anatomical structures of interest - the sternum and costal cartilages.

2.1 Deep learning theory

Deep learning is a subset of machine learning and plays a central role in modern medical image analysis [8]. This section introduces the fundamentals of machine learning and deep learning, including the structure and function of neural networks, and explains the deep learning models nnU-Net and SCN.

2.1.1 Machine learning

Machine learning is a subfield of artificial intelligence focused on developing algorithms that allow computers to learn patterns from data and make predictions or decisions without being explicitly programmed for every task. Contrary to traditional programming, which relies on fixed rules, machine learning methods are built on probabilistic methods that can capture complex relationships and handle uncertainties within data. This makes machine learning effective in fields where data is abundant, but where explicit rule-based solutions are impractical, such as image recognition or natural language processing [9].

2.1.2 Deep learning

Deep learning uses a learning method called artificial neural networks, which processes the data through a series of nodes called neurons. Each neuron has an associated weight that is iteratively learned to minimize prediction errors. These neurons are typically organized into different layers: the input layer, which receives raw data and passes it to subsequent layers; the output layer, which generates the final prediction; and one or more hidden layers in between, which learn complex relationships between the input and output. Networks that contain multiple hidden layers are known as deep neural networks (DNNs), and the process of using DNNs for making

predictions is known as deep learning [10], [11]. A crucial part of designing deep neural networks is the selection of a loss function, also known as a cost function, which quantifies how well the model's predictions align with the true data and is what the model uses to learn different problems [12]. Deep learning is commonly used in medical image processing and can be used for several complex tasks, such as disease detection, semantic segmentation of organs, tumors or cartilage, and localization of normal anatomy [13].

2.1.3 Convolutional neural networks

A convolutional neural network (CNN) is a type of neural network designed for processing input data with grid-like topologies, such as images [12]. CNNs introduce two new types of layers called convolutional and pooling layers. In convolutional layers, a convolution operation is performed by sliding a series of filters, also known as kernels, of specified size and with learnable weights, over an image. Each kernel computes the dot product between the filter's weights and the local region in the input and outputs it to a feature map. This allows the network to efficiently learn local features in the input such as edges, shapes or textures and process them into more abstract representations. An example of a convolution operation can be seen in Figure 2.1. Convolutional operations are usually followed by a non-linear activation function, such as the rectified linear unit, which sets all negative values to zero, and an optional pooling layer [12], [14].

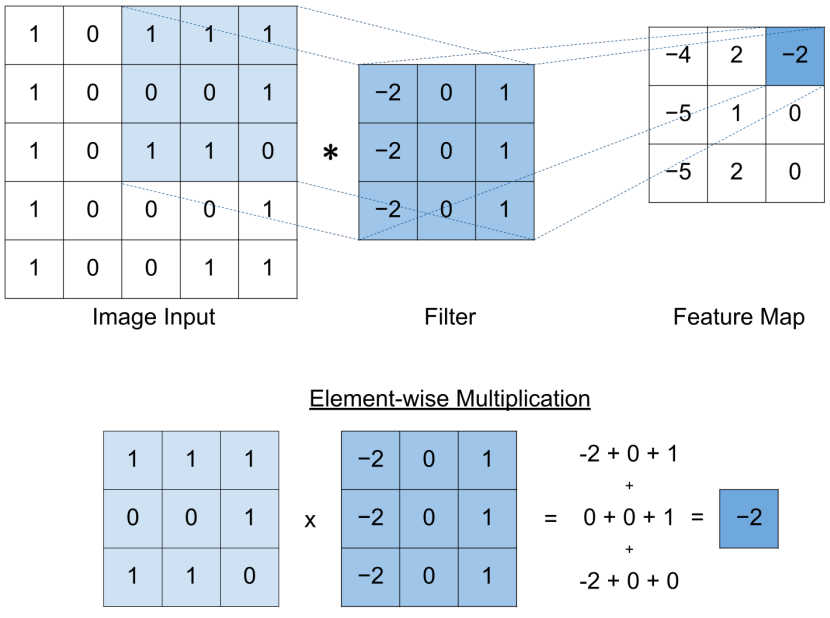


Figure 2.1: Example of a 2D convolution operation with an input image of size 5x5 and a filter of size 3x3.

Pooling layers are used to aggregate information from a neighborhood around each point in the input. This process reduces the dimensionality of the data while introducing approximate invariance to small translations. A commonly used pooling operation is max pooling, which outputs the maximum value within a rectangular

neighborhood. Another widely used method is average pooling, which instead computes the average value within the same region. It is important to note that pooling layers do not contain any learnable parameters; they only summarize information in the input [12], [15]. An example of a max pooling operation with a 2x2 kernel can be seen in Figure 2.2.

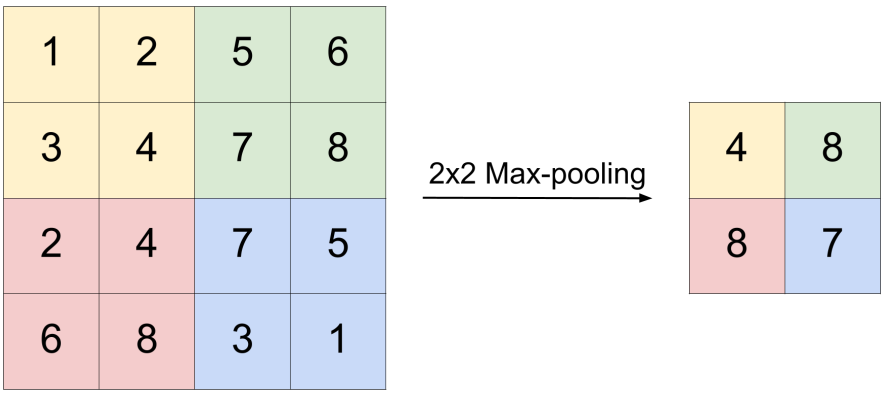


Figure 2.2: Example of a 2D max pooling operation with a 2x2 kernel.

2.1.4 U-Net

A commonly used deep learning architecture for medical image segmentation is a fully convolutional U-shaped network known as the U-Net. The architecture consists of two components: an encoder and a decoder. The encoder, captures contextual information and extracts high-level features in an image by gradually reducing its spatial dimensions and scaling up its feature channels through repeated convolutional and max-pooling layers. The decoder, reconstructs the spatial dimensions while decreasing the number of feature channels using transposed convolutions and concatenates the corresponding feature maps from the encoder via residual connections. This enables the model to combine low-level details with higher-level patterns, allowing for precise localization and accurate segmentations, even with small datasets [16]. An example of a 2D U-Net architecture can be seen in Figure 2.3.

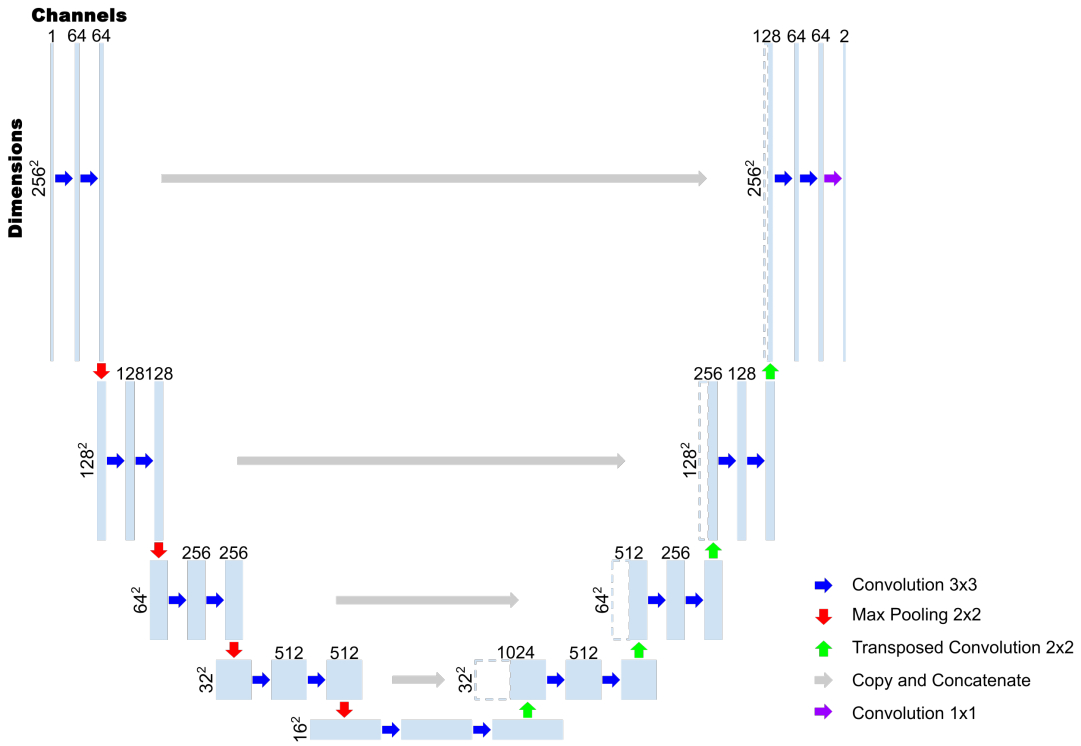


Figure 2.3: Architecture of a 2D U-Net with an input image size of 256x256 pixels and one color channel (greyscale). Output of the network is a two-channel mask corresponding to a binary segmentation with one channel for background and one for foreground.

2.1.5 nnU-Net

The deep learning model nnU-Net is a self-adapting framework tailored for medical image segmentation based on the architecture of the U-Net. It features an automated pipeline that dynamically adapts to a given dataset and hardware. This includes automatic adjustments to pre-processing, network architecture, training procedure, and post-processing steps. The main idea behind nnU-Net is to automate and optimize all the design choices needed for the architecture U-Net to perform as well as possible - and for U-Net to be a reliable benchmark when developing new models [17].

The pre-processing starts by cropping the image to nonzero values to reduce the computational complexity. All images are then resampled to a uniform voxel spacing and finally the image intensities are normalized using different strategies depending on the modality of the images.

The network architecture is configured based on dataset properties and available hardware. This includes determining the input patch size, number of pooling operations, convolutional kernel sizes, and number of feature maps per layer. It creates three different U-Net variants to choose from, depending on the anisotropy and size of the input images: 2D full-resolution, 3D full-resolution, or 3D low-resolution.

The training configuration is also dynamically adapted. This involves setting the

batch size, optimizer parameters, and learning rate schedule. nnU-Net uses a fixed number of training epochs and iterations per epoch, and uses five-fold cross-validation to evaluate performance and prevent overfitting. nnU-Net automatically splits the provided dataset into 80% for training and 20% for validation. Data augmentation is performed in each iteration during training, using a standardized set of geometric and intensity-based transformations [17].

nnU-Net uses a combination of Dice loss and Cross-Entropy loss, adapted based on validation performance. Dice loss is derived from the Dice coefficient, which quantifies the overlap between predicted and ground truth segmentations [18], ranging from 0 (no overlap) to 1 (perfect overlap). The loss decreases as overlap improves and increases with more false positives or false negatives. To support multi-class segmentation, which is common in medical imaging, nnU-Net employs a multi-class variant of Dice loss [17]. In this formulation, the Dice loss ranges from 0 (no overlap) to -1 (perfect overlap), with lower values indicating better segmentation performance

Cross-Entropy loss is commonly used for pixel-wise classification tasks in image segmentation. Instead of evaluating the overlap between entire regions, this loss function compares the predicted class probabilities for each voxel to the ground truth class label. During training, the model calculates a probability distribution over all possible classes for every voxel in the input image. The Cross-Entropy loss then penalizes incorrect predictions, such that the model yields a low loss when it assigns a high probability to the correct class and a higher loss when the prediction is uncertain or incorrect. The lowest value is 0 [19].

By combining Dice loss and Cross-Entropy loss, nnU-Net can train on both regional and voxel information, such that it can learn structures of different sizes as well as converge faster. The total loss is calculated as in Equation 2.1 and has -1 as the lowest possible value.

$$Loss_{Total} = Loss_{Dice} + Loss_{Cross-Entropy} \tag{2.1}$$

Post-processing is applied using connected component analysis. It checks if performance is improved if only the largest connected component is kept for each class in the dataset. This helps eliminate false positives and refine the final segmentation output [17].

2.1.6 SpatialConfiguration-Net

SpatialConfiguration-Net (SCN) is a CNN-based architecture designed for anatomical landmark localization in medical images, particularly effective when training data is limited.

SCN divides the landmark prediction task into two complementary components: one capturing local appearance and the other modeling global spatial configuration. The local appearance component uses heatmap regression, a technique where the network predicts a probability distribution in the form of a heatmap around each landmark, where high values indicate likely positions. The local appearance component aims

to transform the input into locally accurate heatmaps. However, solely relying on local appearance can result in errors where anatomically ambiguous regions cause the model to predict the wrong landmark.

To mitigate this, SCN introduces a spatial configuration component, which aims to capture the anatomical relationship between the different landmarks. The final landmark localization is obtained by an element-wise multiplication of the heatmaps of both the local and the spatial component, combining precise detail with contextual information [20], [21].

The heatmap g_i for each landmark L_i, $i = \{1, ..., N\}$ with coordinate $\overset{*}{x}_i$ and dimensionality d used for heatmap regression is defined as the Gaussian function:

$$g_i(x; \sigma_i) = \frac{\gamma}{(2\pi)^{d/2}\sigma_i^d} exp\left(\frac{\|x - \overset{*}{x}_i\|_2^2}{2\sigma_i^2}\right) \tag{2.2}$$

Here, γ is an intensity scaling factor and σ_i is the standard deviation (spread) of the heatmap for landmark i. In this setup, σ_i is used as a trainable parameter in the network, allowing it to be learned together with the network weights $\mathbf{w}$, and biases $\mathbf{b}$. This allows the model to learn the optimal heatmap size and peak separately for each landmark [21].

The loss function for SCN is formulated as:

$$\underset{\mathbf{x},\mathbf{b},\sigma}{min} \sum_{i=1}^{N} \sum_{x} \|h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}) - g_i(\mathbf{x}; \sigma_i)\|_2^2 + \alpha\|\boldsymbol{\sigma}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 \tag{2.3}$$

Where $h_i(\mathbf{x}; \mathbf{w}, \mathbf{b})$ is the model's heatmap predictions, α defines the scaling factor for the penalty term from $\boldsymbol{\sigma}$, and λ is the scaling factor for the penalty term for $\mathbf{w}$. The first term has the trivial solution $\boldsymbol{\sigma} \to \infty$, where infinitely large heatmaps cover the whole volume. To avoid this, the penalty term for $\boldsymbol{\sigma}$ is added, which prefers small values. This trade-off enables the model to predict narrow heatmaps for high-confidence landmarks and wider heatmaps for landmarks that are more challenging to predict [21].

The evaluation criteria for selecting the best performing model is the inter-point error (IPE), which is defined as the Euclidean distance between a target point $\overset{*}{x}_i$ and a predicted point $\hat{x}_i$:

$$IPE_i = \|\overset{*}{x}_i - \hat{x}_i\|_2 \tag{2.4}$$

Another evaluation criterion used to describe the model's performance is the proportion of outliers O_r, which is defined as the proportion of predicted points that lie outside a radius r from their respective target points.

2.2 Medical data

The input data is 3D CT images from NMDID, manual segmentations and manual landmarks. A CT image is a greyscale reconstruction, commonly used to reconstruct

the human body. The grey tones are the visualization of the Hounsfield Unit (HU) - how much the tissue absorbs the x-ray beam. High values represent high-density tissue, such as bone, and are visualized in bright shades. Low values represent low-density tissue and are visualized in dark shades [22].

The CT images are obtained in the Digital Imaging and Communications in Medicine (DICOM) file format, which is the international standard for medical images [23]. The DICOM format was developed to ease the storage and exchange of medical data since different medical equipment manufacturers used different data formats. DICOM holds image data in several layers (slices). One slice is a 2D image and several slices together makes a 3D image. DICOM also contains other information that is of importance to describe the image, such as the object's description and the patient's data. This information is kept in a header [24]. Additionally, the header contains information about the image's position, orientation, and coordinate system. Position is defined by the coordinates of the upper-left corner of the image, while orientation specifies the direction of the first row and column relative to the patient. DICOM images use the LPS (Left, Posterior, Superior) coordinate system, where axes increase from right to left, anterior to posterior (front to back), and inferior to superior (bottom to top) [25]. LPS is a right-handed coordinate system.

The manual segmentations are obtained in the Nearly Raw Raster Data (NRRD) file format. NRRD, like DICOM, is suitable for medical images. Like DICOM, it saves image data and other necessary information in a header. Instead of keeping the image data divided in slices it is saved in the same file for both 2D and 3D images. The image data and header can be either one united file or two separate ones [26]. The NRRD header, like the DICOM header, contains information about the image's position, orientation, and coordinate system. Additionally, it specifies the spacial dimension of the image. In NRRD, *position* refers to the location of the first voxel in space, and *direction* is described by vectors that define the orientation and spacing of each axis. Unlike DICOM, which uses a fixed LPS coordinate system, NRRD supports multiple coordinate systems. A commonly used one is RAS (Right, Anterior, Superior), where the axes increase from left to right, back to front, and bottom to top [27]. RAS is also a right-handed coordinate system, so conversions between LPS and RAS do not require mirroring, but do involve flipping the direction of certain axes, along with rotation and translation.

The manual landmarks are stored in Comma-Separated Values (CSV) format [28]. Each CSV file includes a header row that specifies the names of the landmarks and the coordinate axes (X, Y, Z). Subsequent rows contain the corresponding X, Y, and Z coordinates for each landmark.

2.3 Anatomy

Knowledge of the sternum and costal cartilages is essential for performing and evaluating segmentation and landmarking of these anatomical features. These structures are colored in Figure 2.4: the sternum consists of three parts - manubrium (green), sternal body (orange), and xiphoid process (yellow), and the costal cartilage (blue).

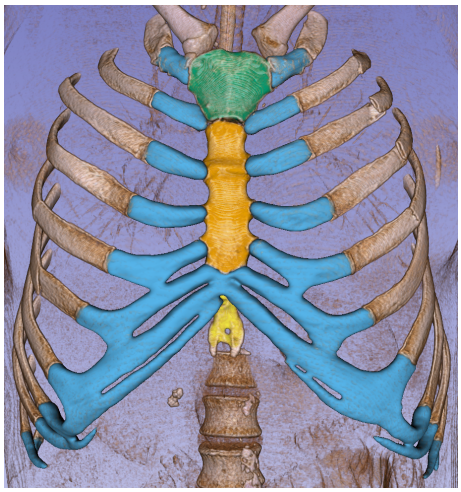
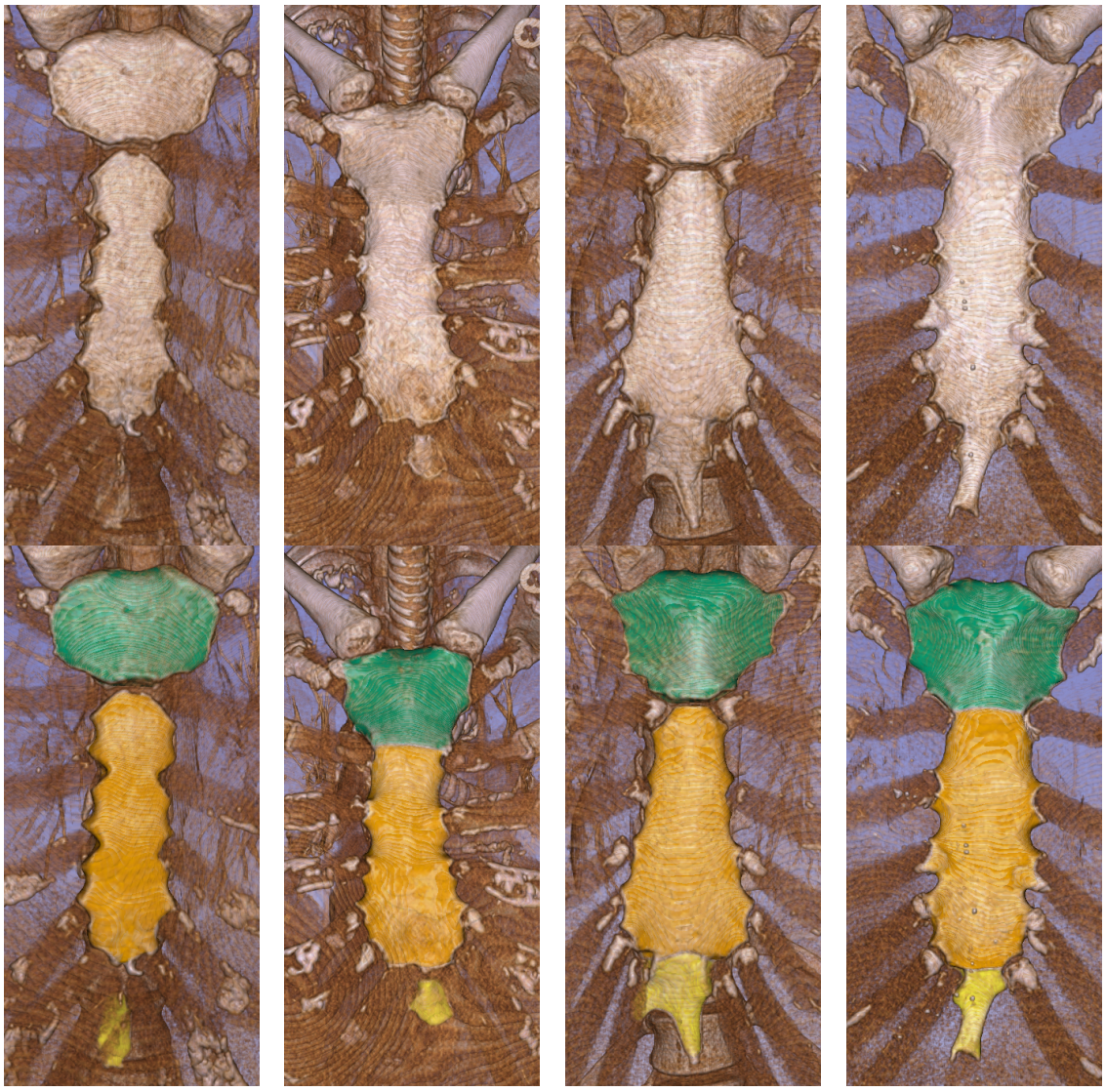


Figure 2.4: CT image of thoracic region with sternum and costal cartilage segmented.

2.3.1 Anatomy of the sternum

The sternum, located at the center of the ribcage, has an important role in protecting the thoracic organs by forming a rigid structure together with the ribs and costal cartilages. It is composed of three segments: the manubrium, the sternal body and the xiphoid process. The sternal body is connected to the manubrium at the manubriosternal joint and the xiphoid process at the xiphisternal joint [29]. The joints vary anatomically between individuals and depending on age. They can be separated, partially fused, or fully fused, either from birth or as a result of physiological changes over time [6]. The basic structure of the sternum and its joints is illustrated in Figure 2.5, and examples of anatomical variation are shown in Figure 2.6.

Figure 2.5: Sternum divided into manubrium (green), sternal body (orange), and xiphoid (yellow). Manubriosternal- and xiphisternal joints are circled.



(a) Manubrium, sternal body and xiphoid divided in three parts

(b) Manubrium and sternal body fused at the manubriosternal joint

(c) Sternal body and xiphoid fused at the xiphisternal joint

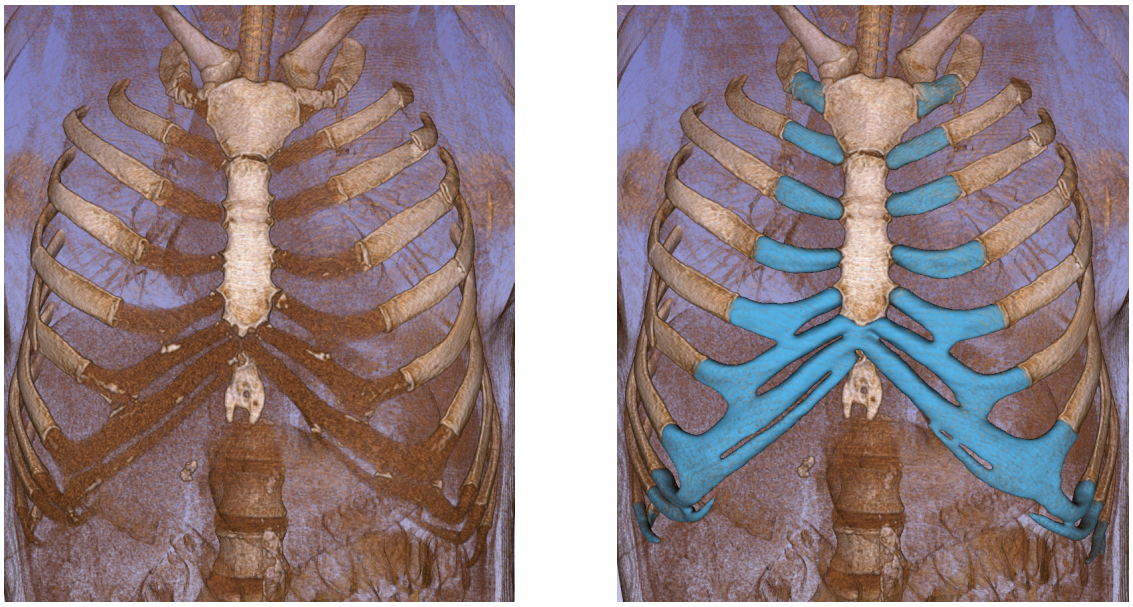
(d) Manubrium, sternal body and xiphoid fused in one part

Figure 2.6: Four different sternums with different types of compositions.

2.3.2 Anatomy of the costal cartilages

The costal cartilages are extensions of the ribs that attach either to the sternum or to another cartilage. Some ribs have costal cartilages that do not attach to anything, these are referred to as floating ribs. Traditionally, the seven uppermost costal cartilages are considered to attach directly to the sternum, while the next three attach to the cartilage above them or remain unattached, thus also being classified as floating. Below these ten ribs, there are two additional ribs that do not have costal cartilage. These are excluded from the scope of this thesis [29].

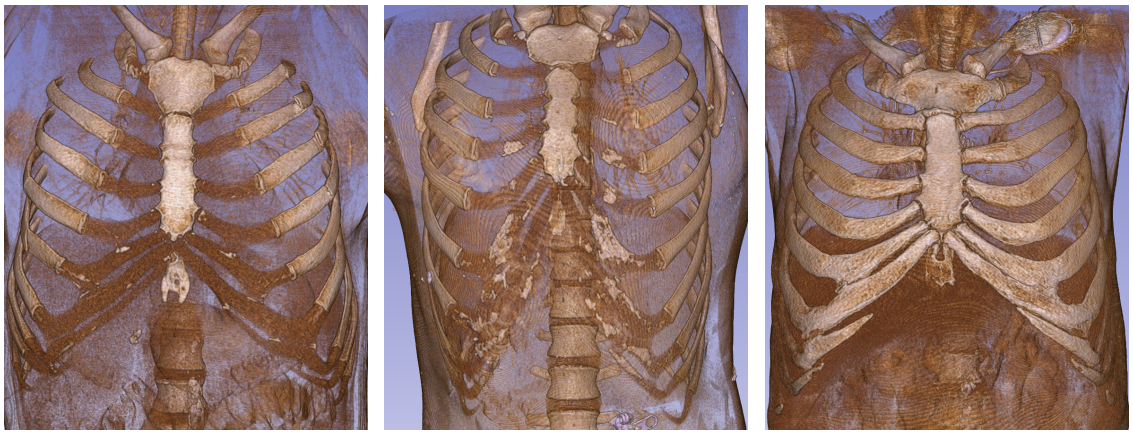
In Figure 2.7, the points of attachment for the costal cartilages can be seen. The first cartilage attaches to the manubrium, while the second attaches at the junction between the manubrium and the sternal body, at the level of the manubriosternal joint. The third to seventh costal cartilages typically attach to the sternal body, while the eighth to tenth usually attach to the cartilage above. In some individuals, the tenth may be unattached and considered floating, though this varies.



(a) CT image of thoracic region **(b)** Costal cartilage segmented

Figure 2.7: Structure for costal cartilage

Costal cartilages are cartilaginous structures and are softer than the bony sternum. Due to their lower density, they appear less bright than the sternum on CT images. The visual difference can be seen in the volume in Figure 2.7a. It is however common for calcification to appear in the cartilage. This increases with age and some even develop fully calcified cartilages. Three different degrees of calcified cartilages is shown in Figure 2.8. The calcification is more stiff than cartilage and display similar structure as bone [30].



(a) Minimal calcification **(b)** Moderate calcification **(c)** Extensive calcification

Figure 2.8: Three different degrees of calcification in the costal cartilages

3

Methods

This chapter describes the methodology used in the study, including the input data, hardware setup, software tools, and the deep learning models nnU-Net and SCN. It explains how the datasets were generated and formatted, as well as how the models were trained and evaluated.

3.1 Input data

The data used in this thesis consists of CT images of human torsos from the NMDID dataset, referred to here as volume data. The volumes were provided in DICOM format, while manual segmentations of the sternum and costal cartilages were supplied as NRRD files. In total, six segmentations were available.

Landmark data were provided as 20 CSV files containing the physical coordinates of over 200 points, of which 34 were anatomical landmarks. Anatomical landmarks are well-defined, anatomically meaningful points on biological structures that can be consistently identified across individuals, providing a reliable basis for comparative analysis [31]. Only anatomical landmarks have been studied further.

3.2 Hardware and software

This section presents the hardware and software used in this study.

3.2.1 Hardware

All data pre-processing and model training were performed on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM), an Intel Core Ultra 7 265KF CPU (20 cores, 20 threads), and 64 GB of system RAM.

3.2.2 3D Slicer

The open-source tool *3D Slicer* was used to visualize CT volumes and segmentations, as well as to perform manual segmentations [32], [33]. 3D Slicer uses the voxel structure of CT images to enable visualization of slices in multiple anatomical planes: axial (horizontal), sagittal (side), and coronal (front). In addition to these

2D views, both the volume and segmentations can be rendered in 3D using the *Volume Rendering* module. The 3D visualization can be adjusted by setting intensity thresholds, allowing certain structures to be highlighted or hidden. For instance, thresholds were tuned to suppress the visualization of skin and other obscuring tissues, while preserving visibility of bones and cartilage. These various views served as the basis for manual segmentation.

Manual segmentation was performed using the *Segment Editor* module in 3D Slicer [34]. This tool allows editing in both 2D and 3D views and provides several methods for segmentation. The methods primarily used in this work were *Paint*, *Draw*, and *Erase*. Additionally, the *Segmentations* module was used to adjust the opacity of the segmented regions, enabling the underlying anatomical structures, specifically the sternum and costal cartilage, to remain visible through the segmentation.

3.2.3 Programming language

All programming in the thesis was conducted using Python 3.13. For data preprocessing, NumPy [35] was used for numerical operations and array handling, while SimpleITK [36] was used for image loading, resampling, and spatial transformations. Deep learning models were implemented using PyTorch [37] with data augmentation from TorchIO [38]. Additionally, segmentation data generated in 3D Slicer was modified and accessed using the slicerio Python package.

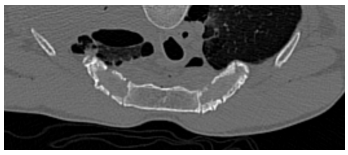
3.3 Segmentation of sternum and costal cartilage

To segment the sternum and costal cartilages from CT images, the deep learning model nnU-Net was used. It was selected due to its strong performance in medical image segmentation and its ability to automatically adapt pre-processing, architecture configuration, and training to a specific dataset [17]. This section describes the full segmentation process, including dataset creation, data preparation and formatting, use of nnU-Net, and evaluation of the results.

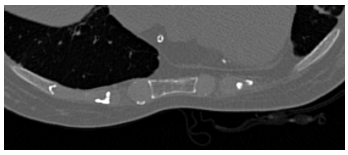
3.3.1 Expanding dataset

To ensure adequate dataset size for the deep learning model, the initial set of six segmentations was expanded to ten through additional manual segmentation in 3D Slicer, using the original segmentations as guidance. When selecting volumes for segmentation, anatomical outliers were excluded, such as volumes with fully calcified costal cartilages or visible signs of surgery affecting the sternum or costal cartilages, as these cases do not represent the general population.

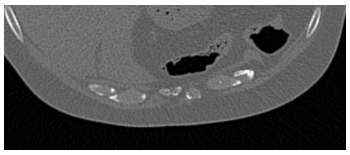
Despite the exclusion of anatomical outliers, there remained considerable variation in image quality, visibility, and anatomical structure across the selected volumes. These differences influenced the manual segmentation process and consequently affected the overall quality and consistency of the training set. Figure 3.1 illustrates this variability by showing axial slices from three different volumes. Both image quality and anatomical structure differ noticeably between Figure 3.1a, b, and c.



(a) View for segmentation of upper costal cartilage and manubrium



(b) View for segmentation of middle costal cartilages and sternal body



(c) View for segmentation of lower costal cartilages and xiphoid process

Figure 3.1: Axial view used for segmentation, shows three different slices from three different volumes.

The segmentation included four anatomical labels: manubrium, sternal body, xiphoid process, and costal cartilage. To ensure uniformity across the dataset, a naming and labeling standard was established, using the labels *Manubrium*, *Sternal_body*, *Xiphoid*, and *Cartilage* with corresponding label values of 1, 2, 3, and 4. To have a uniform dataset is vital for nnU-Net's, and other machine learning models, ability to learn patterns.

Segmentations were saved as NRRD files and named after the corresponding volume. Because segmenting the sternum and the ten upper costal cartilages is highly time-consuming, only the sternum and the top five costal cartilages were segmented in four volumes. These four segmentations, combined with six pre-existing ones, formed an initial dataset to evaluate whether the selected model, nnU-Net, was suitable for this specific segmentation task.

3.3.2 Initial preparation and formatting

To enable model training, the dataset needed to be processed to ensure consistency and correct alignment in the dataset. The dataset also needed to be compatible with nnU-Net. To be trainable with nnU-Net, a dataset must follow a specific structure: it should include a folder of raw CT images, a folder of corresponding segmentation masks, and a `dataset.json` file specifying key metadata such as label values and the number of training samples. All components must follow a naming convention that links each volume with its corresponding segmentation, ensuring compatibility with nnU-Net's data loader [17]. This preparation was carried out in four steps, each implemented as a separate Python script:

1. Verification and standardization of label values: Although a standardized naming scheme was followed during segmentation in 3D Slicer, the corresponding numerical label values were initially found to be inconsistent. To address this issue and to ensure consistency across all segmentations, a pre-processing step was introduced to standardize the label values according to a predetermined mapping between labels and numerical values.
2. DICOM to NRRD conversion: The CT scans were originally provided in DICOM format and needed to be converted to NRRD to match the format of the segmentations. For this purpose, voxel data along with relevant metadata,

such as image origin, direction, and spacing, was extracted and converted using the SimpleITK Python package.

3. Cropping of volumes and segmentations: To reduce memory usage during pre-processing and reduce computational cost during inference (prediction phase), all CT volumes and their corresponding segmentations were cropped to a standardized bounding box. This bounding box was determined by computing the minimum and maximum nonzero indices along each axis across all segmentations in the dataset, then adding a fixed padding margin in all directions. All volumes and segmentations were cropped to this region, ensuring that relevant anatomical structures were retained while excluding irrelevant background.

4. Renaming and saving to nnU-Net format: To prepare the dataset for training with nnU-Net, all CT volumes and corresponding segmentations were renamed to comply with nnU-Net's standardized naming convention. The renamed volumes and segmentations were then saved to the designated training and label directories.

3.3.3 nnU-Net training details

After the initial dataset preparation and formatting, pre-processing was carried out using nnU-Net's built-in plan and pre-process function with default settings. This step generated three configurations - 2D full-resolution, 3D full-resolution, and 3D low-resolution - and performed cropping, resampling, and normalization on the image volumes. The pre-processing aimed to reduce computational complexity and standardize the input data for training.

For training, the 3D full-resolution configuration was selected, as the task involved segmenting 3D medical images, with the goal of matching or exceeding the quality of manual segmentations. Training was performed using nnU-Net's default settings and five-fold cross-validation. The default settings were maintained because nnU-Net is designed to automatically optimize architectural and training parameters, and previous studies have demonstrated that the default configuration performs reliably across various datasets. Early training confirmed that the default settings gave reasonable results.

Once training was complete, segmentations were predicted and post-processed using nnU-Net's corresponding built-in functions. Post-processing was applied to remove false positives that were incorrectly identified during prediction. Training the model took approximately 87.5 hours in total (17.5 hours per fold across five folds), and generating a segmentation, including post-processing, took about ten minutes per volume.

The model used was the second version of nnU-Net, nnU-Net v2, and with an 8 GB VRAM limit [39]. During testing, VRAM limits of up to 24 GB were tried. During testing, higher VRAM limits of up to 24 GB were explored, but no significant differences in performance were observed. Therefore, an 8 GB limit was chosen to reduce training time.

3.3.4 Expanding to final dataset

After training nnU-Net on a dataset containing segmentations of the sternum and the top five costal cartilages, the initial predictions showed promising results. This motivated an expansion of the dataset to include all ten costal cartilages. The segmentations were then updated accordingly.

Several weaknesses were identified when evaluating the results of the training on the five topmost cartilages, particularly in the areas where the costal cartilages attach to the ribs. In response to these deficiencies, a more focused effort was made to refine and improve the manual segmentations in these specific regions, ultimately improving the quality of the dataset for further training. Segmenting a single volume took approximately 30 hours, depending on the quality of the CT scan and the anatomical complexity.

To increase the efficiency of the manual segmentation process, a semi-automatic approach was adopted. The first step involved manually segmenting the sternum and the ten topmost costal cartilages in five volumes using 3D Slicer. These initial segmentations were then used to train a model. After training, the model predicted segmentations for additional volumes, yielding nearly complete results. These preliminary outputs were manually refined in 3D Slicer to create a training set of consistent quality. In addition to training, prediction, and post-processing time, manual refinement typically took around 30 minutes per volume. This semi-automatic approach improved efficiency and enhanced consistency by reducing the manual workload and minimizing human error.

This process was repeated to investigate the impact of dataset size on the results, continuing until a sufficiently large dataset was obtained, defined as one where the model's segmentation errors were comparable to those made by humans. Ultimately, datasets containing five, eight, and ten segmentations were evaluated.

3.3.5 Evaluation

Once the training was done, the model's predictions were evaluated in different aspects — Dice coefficient, visual inspection and in comparison with manual segmentation. The model was evaluated for each different size of the training set.

3.3.5.1 Dice coefficient and visual inspection

Both the total Dice coefficient and the Dice coefficient for each class were recorded and analyzed. The total Dice coefficient provided an overall measure of the model's performance, while the class-specific Dice coefficients made it possible to see if the model had difficulties predicting certain classes. This helped in understanding how individual class performance contributed to the total performance.

In addition to evaluating the Dice coefficients, the predicted segmentations were visually reviewed using 3D Slicer. Each prediction was compared to the corresponding CT volume to assess how well it aligned with the actual anatomical structures. This visual inspection made it possible to identify areas where the model consis-

tently performed well or poorly. While the Dice coefficients give a general idea of how the model performed across different classes, they do not provide information about specific regions - something the visual analysis helped to highlight. The visual inspection was conducted on volumes that were not part of the validation set, providing a broader perspective for the evaluation.

3.3.5.2 Model prediction vs. manual segmentation

To evaluate whether the model's predictions were comparable or even better than manual segmentations, one model prediction was compared to two manual segmentations of the same volume, done by two different individuals. The manual segmentations were also used to compare the work of two different individuals, to get an idea of the influence of the human factor.

To achieve unbiased manual segmentations, the volume that was segmented was chosen by a person who would not perform manual segmentation. The volume chosen had a CT image, which was said to have a medium difficulty level, based on image quality and the amount of calcified cartilage. Furthermore, the model's prediction was completed after the manual segmentations were done, so that the manual segmentations would not be affected by the model's prediction. To reduce the time spent on segmenting, only the five topmost costal cartilage were segmented.

3.4 Landmarking of manubrium and sternal body

To localize anatomical landmarks on the manubrium and sternal body, the CNN-based SpatialConfiguration-Net (SCN) was used. SCN was chosen for its ability to model both local image appearance and global anatomical relationships, making it well-suited for medical landmarking tasks with limited training data [20], [21]. This section outlines the full pipeline used for landmark localization in this thesis: pre-processing of segmentation and landmark data, model architecture and training setup, and the post-processing steps used to extract predicted coordinates from the model outputs.

3.4.1 Pre-processing

The starting point for the landmarking task was a set of segmentations generated in the previous step. The segmentations were in NRRD format, covering the manubrium, sternal body, xiphoid process, and costal cartilage, along with corresponding CSV files containing the landmarks in physical coordinates. The landmarking task was restricted to predicting landmarks around the manubrium and sternal body, as these structures exhibit less anatomical variation between individuals compared to the xiphoid process and costal cartilage. On account of this, the segmentations for these two classes were excluded, and only the first 34 lines of each landmark file, corresponding to the selected anatomical landmarks, were retained. Furthermore, the label values of the manubrium and sternal body were unified into a single class to produce a binary segmentation mask. This simplification was made because distinguishing between the two structures was no longer necessary. Finally,

all segmentations were resampled to the mean resolution of the dataset and padded to uniform dimensions to ensure standardized inputs to the model.

For each landmark file, the anatomical landmarks were extracted, and temporary Gaussian heatmaps were generated around the corresponding points. These heatmaps were stored in a 4D volume of shape (N, D, H, W), where N was the number of landmarks and (D, H, W) were the dimensions of the pre-processed segmentation. While these heatmaps were not used during training, they served an important role in handling data augmentation: specifically, in estimating the updated landmark coordinates after applying spatial transformations in the model's data loader.

Directly transforming 3D landmark coordinates to match the augmentations applied to the images proved non-trivial. As a workaround, the Gaussian heatmaps were augmented using the same transformation parameters as the input images. The new landmark positions were then estimated by identifying the voxel with the maximum intensity in each transformed heatmap.

While this method worked in practice, it had some notable drawbacks. It was considerably slower than transforming coordinates directly, as it required augmenting sets of 3D heatmap volumes rather than a small set of points. Moreover, applying transformations to the heatmaps introduced a risk of misalignment due to interpolation artifacts during augmentation, potentially shifting the ground truth voxel away from its intended location.

3.4.2 Model architecture

The network was built using PyTorch [37]. It was initialized as described in [21], with 3x3x3 kernels for the convolutions in the local appearance component and 7x7x7 kernels in the spatial configuration component. Resampling was done using trilinear interpolation. Dropout [40] of 0.5 was included after the first convolution in each layer to improve the generalization of the model.

The local appearance component began with an initial convolutional layer that increased the number of channels from 34 to 64. This was followed by three stages, each consisting of two consecutive convolutional layers followed by an average pooling layer to reduce the spatial resolution.

To preserve information across layers, residual connections were used. These included an upsampling layer in each set to match the spatial dimensions before adding the residual. Finally, a convolutional output layer reduced the number of channels back to 34, producing the local appearance heatmaps.

The spatial configuration component processed the heatmaps generated by the local appearance module to model spatial relationships between landmarks. It began with a downsampling layer that rescaled the resolution by 1/4th, and an initial convolutional layer that increased the number of channels to 64.

This was followed by two more convolutions and one last convolutional layer with a tanh activation function, decreasing the channels back to 34 and rescaling the

output to lie between -1 and 1. Finally, the outputs were rescaled back to the original resolution using an upsampling layer, producing the spatial configuration heatmaps.

The local appearance and the spatial configuration heatmaps were multiplied element-wise to produce the final set of predicted heatmaps. To train the model using heatmaps of varying size for each point, the standard deviation of the heatmaps, $\boldsymbol{\sigma}$, was added as a trainable parameter in the network.

3.4.3 Experimental setup and training details

The standard deviation σ_i for each landmark was initialized to 3.0, which resulted in heatmaps with sufficient spread for the model to learn the spatial relationships. The scaling factor γ for the heatmaps was set to 1000, while the regularization terms α and λ were set to 100 and 0.0005, respectively. These values provided a good balance between the different loss components, providing numerical stability during training.

The loss function was minimized using the Adam optimizer with an initial learning rate of 0.0001, where $\beta_1 = 0.5$ and $\beta_2 = 0.999$ control the decay rates of the moving averages of the gradients and their squared values, respectively. The model was trained for a fixed number of 5000 epochs with a linearly decaying learning rate after epoch 2500. These hyperparameters and learning rate schedule led to stable convergence within the allotted training time.

Data augmentation was performed on-the-fly using torchio [38]. The images and landmarks were randomly translated by [-5, 5] pixels and rotated by [-5, 5] degrees, in each dimension, as well as randomly scaled by [0.9, 1.1].

During training, performance was evaluated on a separate validation set using the average IPE as the selection criterion. The model weights corresponding to the epoch with the lowest validation IPE were saved and used for final evaluation on the test set.

3.4.4 Post-processing

The voxel with the highest intensity in each heatmap was identified and treated as the predicted landmark location to obtain the predicted physical coordinates from the model output. These voxel indices were then converted to physical coordinates using the corresponding input segmentation's origin, spacing, and direction matrix.

4

Results

This chapter presents the results of automatic segmentation and landmarking across different dataset sizes. The segmentation part includes the progression from proof of concept to final evaluation, including a comparison to two manual segmentations. The landmarking results focus on the final evaluation.

4.1 Segmentation results

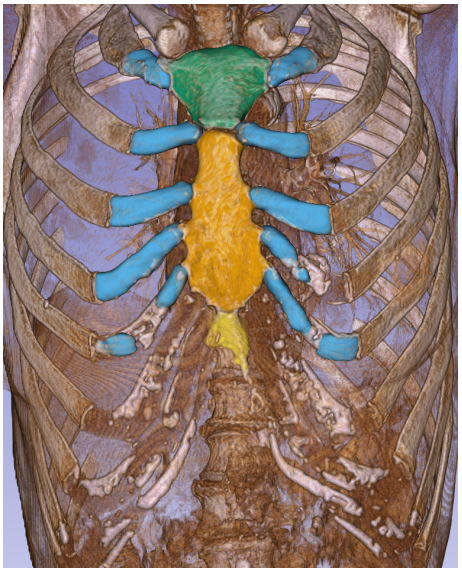
This section presents the results from applying nnU-Net to the segmentation task. It begins with a proof of concept to evaluate whether nnU-Net is suitable for segmenting the sternum and costal cartilages. It then details the final evaluation, where the model was trained on datasets of five, eight, and ten segmentations, respectively. Finally, it compares two manual segmentations and one model prediction to examine the model's performance and explore the accuracy of the manually created dataset.

4.1.1 Proof of concept: testing nnU-Net for segmentation

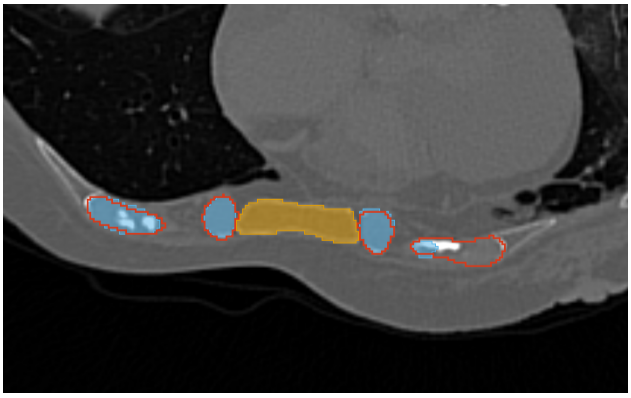
The experiment using a dataset limited to the five topmost costal cartilages produced promising results. After training on a dataset consisting of ten segmentations, nnU-Net successfully segmented the manubrium and sternal body but struggled with the xiphoid process and costal cartilages. The segmentation of the costal cartilages showed two main issues: incomplete attachment into the ribs and incorrect prediction of the number of cartilages. The model often failed to capture the full extent of the cartilage attachments, resulting in cartilages that were too short. Additionally, nnU-Net frequently predicted an incorrect number of cartilages and showed some difficulty in handling calcified regions. Despite these limitations, the results suggested that nnU-Net was a suitable model for this task, as it was able to segment the manubrium, the sternal body and much of the costal cartilage reasonably well, even with a minimal dataset.

Figure 4.1b shows a model prediction of the manubrium, sternal body, xiphoid process, and the upper costal cartilages of the volume shown in Figure 4.1a. While the model correctly predicted the number of costal cartilages, it struggled with the segmentation of the lower cartilages and their attachment to the ribs. In Figure 4.1c, the red overlay highlights the ground truth, revealing areas where the prediction deviates from the manual segmentation. The model also showed reduced accuracy in regions with calcification, which affected the segmentation performance.

(a) Volume without segmentation



(b) 3D view of the segmentation



(c) Slice in the axial plane showing missed costal cartilage in the attachment to the rib on the right side of the image, the red circles the ground truth.

Figure 4.1: Model prediction after training on dataset with five costal cartilages segmented in different views.

4.1.2 Final evaluation: segmentation results on complete data

After increasing number of segmented costal cartilages as well as improving the segmentations the results were improved. The problems that nnU-Net struggled with previously, such as the attachment of costal cartilage into the ribs and predictions of the right number of costal cartilages, decreased. In total, the model was used to predict approximately 200 segmentations.

The size of the dataset had a noticeable impact on segmentation performance, as evaluated both visually and quantitatively using the Dice coefficient. Table 4.1 and

Table 4.2 show the Dice coefficients before and after post-processing for each dataset size and each class as well as the total. The results indicate that the segmentation of the manubrium improves with a larger dataset, while the performance for the sternal body, xiphoid process, and costal cartilage remains relatively consistent across dataset sizes. Among the classes, the sternal body achieved the highest mean Dice coefficient, whereas the xiphoid process had the lowest. The xiphoid process was the only structure that consistently benefited from post-processing, with the exception of the sternal body in the five-sample dataset. The total Dice coefficient in the final column shows an overall improvement in segmentation performance with increasing dataset size.

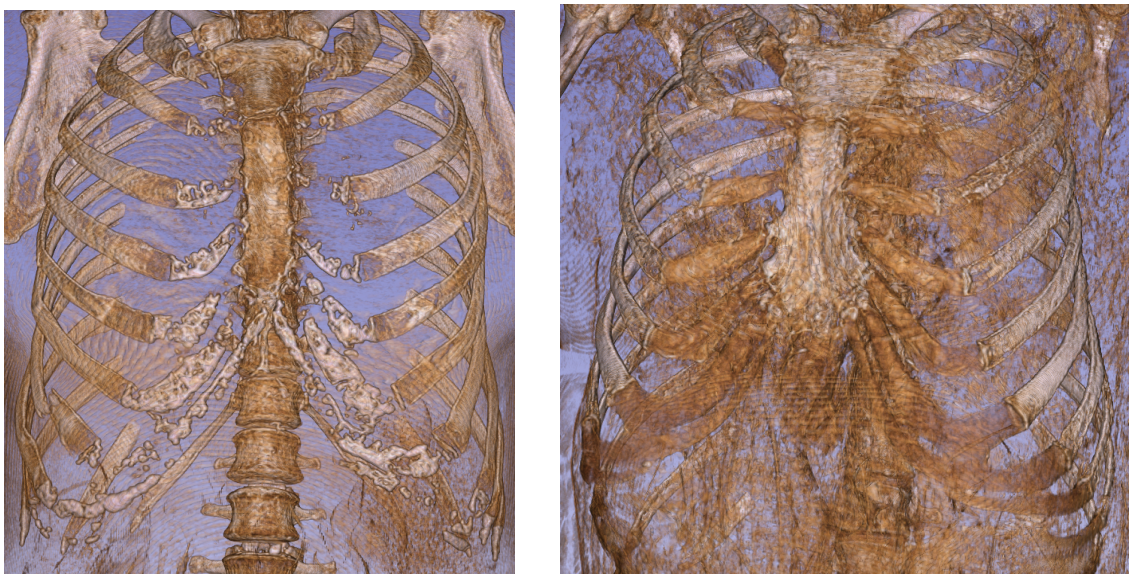
Table 4.1: Dice coefficients for each class and total across different training set sizes after prediction.

Dataset size	Manubrium	Sternal body	Xiphoid process	Costal cartilage	**Total**
5	0.81793	0.94095	0.81825	0.90789	**0.87126**
8	0.91401	0.95923	0.81834	0.91113	**0.90068**
10	0.96433	0.95431	0.81990	0.90605	**0.91115**

Table 4.2: Dice coefficients for each class and total across different training set sizes after post-processing.

Dataset size	Manubrium	Sternal body	Xiphoid process	Costal cartilage	**Total**
5	0.81793	0.94112	0.81952	0.90789	**0.87161**
8	0.91401	0.95923	0.82799	0.91113	**0.90309**
10	0.96433	0.95431	0.83140	0.90605	**0.91402**

To further assess performance, two volumes not included in the validation set were selected for prediction and visual inspection (Figure 4.2). Volume A includes some calcified costal cartilages, while volume B has a grainy appearance. These volumes were chosen to evaluate the model's ability to handle different image characteristics and to examine whether the choice of volumes for the datasets influenced the results. Despite their differences, both volumes were considered representative of the overall volumes, as neither is a clear outlier.



(a) Volume A **(b)** Volume B

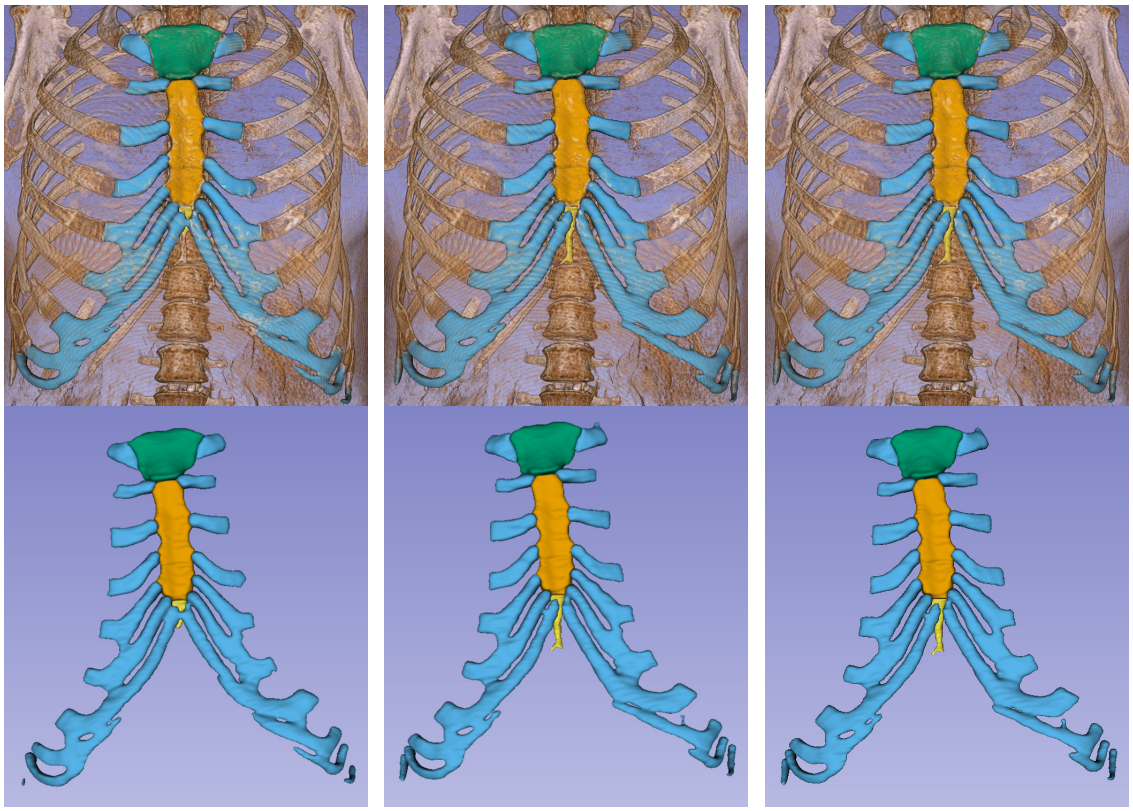
Figure 4.2: Two volumes with different anatomical and image characteristics selected to evaluate the model's performance.

The predictions for volume A across different dataset sizes are presented in Figure 4.3 and Figure 4.4. The analysis begins with the model trained on the five-sample dataset. As shown in Figure 4.3a, the model appears to have managed to segment the manubrium and sternal body. However, a closer inspection of Figure 4.4e reveals that the manubrium was not fully segmented, while the sternal body appears to be accurately segmented in Figure 4.4f.

The model trained on the five-sample dataset also struggled with segmenting the xiphoid process and some of the calcified costal cartilages. A large portion of the xiphoid process is missing in the 3D view (Figure 4.3a), and although it appears in the axial slice (Figure 4.4g), a small segment is still absent. The costal cartilage is partially segmented, with missing sections visible in Figure 4.4h.

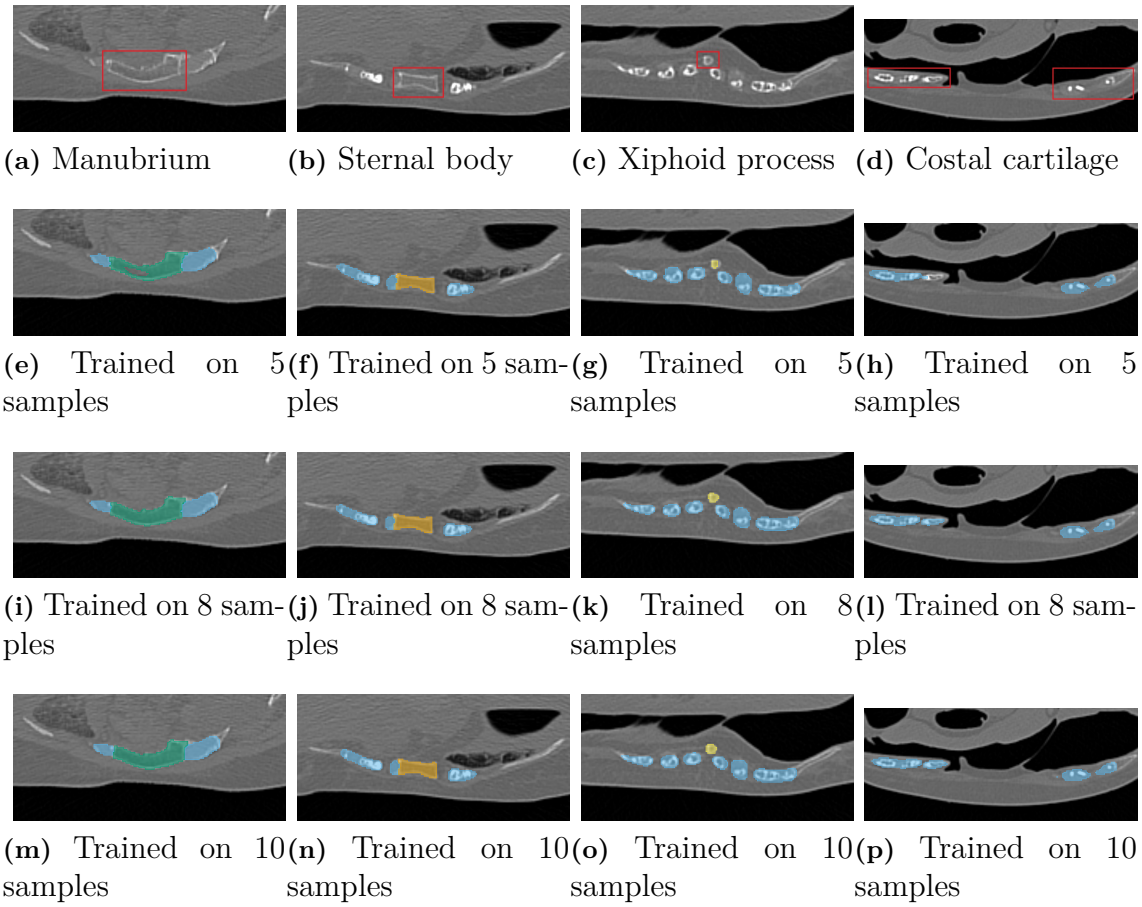
Subsequently, the model trained on the eight-sample dataset shows improved performance, successfully segmenting all anatomical classes (Figure 4.3b). The corresponding axial views, Figure 4.4i-l, further support the observation that the model accurately segmented the volume across all classes.

Similarly, the model trained on the ten-sample dataset also demonstrates strong performance, with all classes segmented seen in Figure 4.3c and Figure 4.4m-p. When comparing the eight- and ten-sample models, no visible differences in segmentation quality are observed for volume A.



(a) Trained on 5 samples **(b)** Trained on 8 samples **(c)** Trained on 10 samples

Figure 4.3: Predicted segmentations of volume A by models trained on datasets with 5, 8, and 10 samples.



(a) Manubrium **(b)** Sternal body **(c)** Xiphoid process **(d)** Costal cartilage

(e) Trained on 5 samples **(f)** Trained on 5 samples **(g)** Trained on 5 samples **(h)** Trained on 5 samples

(i) Trained on 8 samples **(j)** Trained on 8 samples **(k)** Trained on 8 samples **(l)** Trained on 8 samples

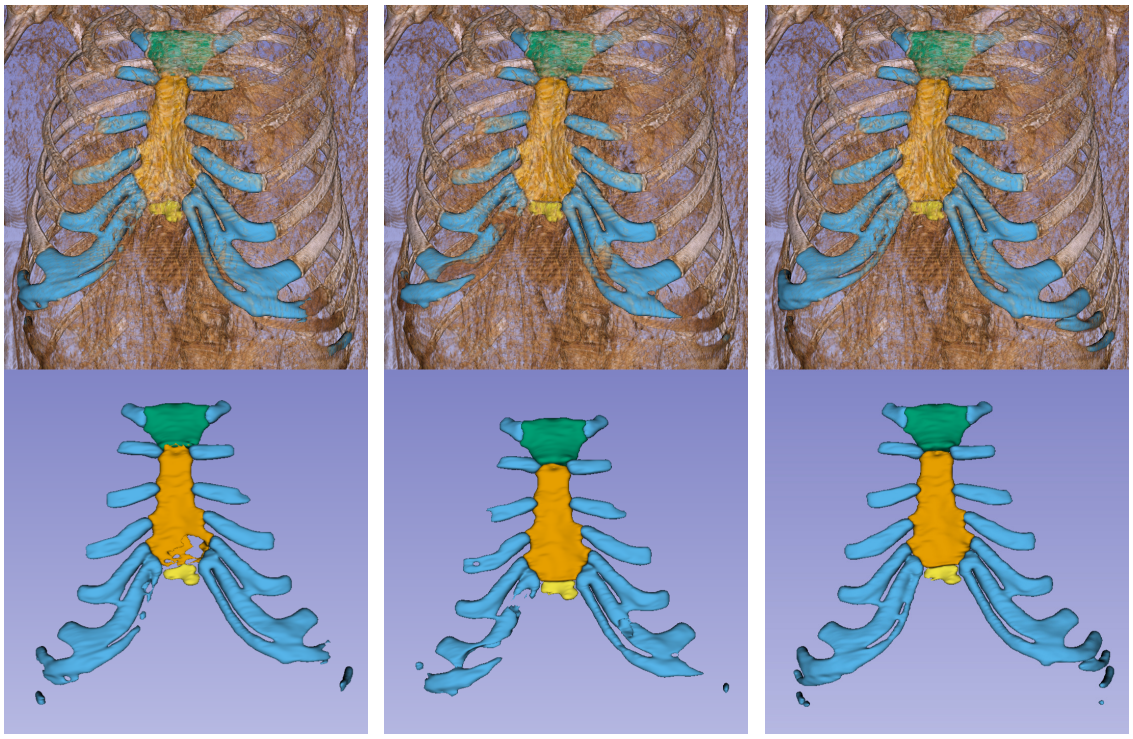
(m) Trained on 10 samples **(n)** Trained on 10 samples **(o)** Trained on 10 samples **(p)** Trained on 10 samples

Figure 4.4: Axial view of different areas of the segmented volume, volume A.

The predictions for volume B with different training set sizes are presented in Figure 4.5 and Figure 4.6. The analysis begins with the model trained on the five-sample dataset. As shown in Figure 4.5a and Figure 4.6e, the model successfully segmented the manubrium. However, this was the only class it segmented well. Figure 4.6f shows that large portions of the sternal body were missed. Additionally, the xiphoid process is only partially segmented (Figure 4.6g) and in the left costal cartilage (Figure 4.6h), which also appears hollowed in the same slice.

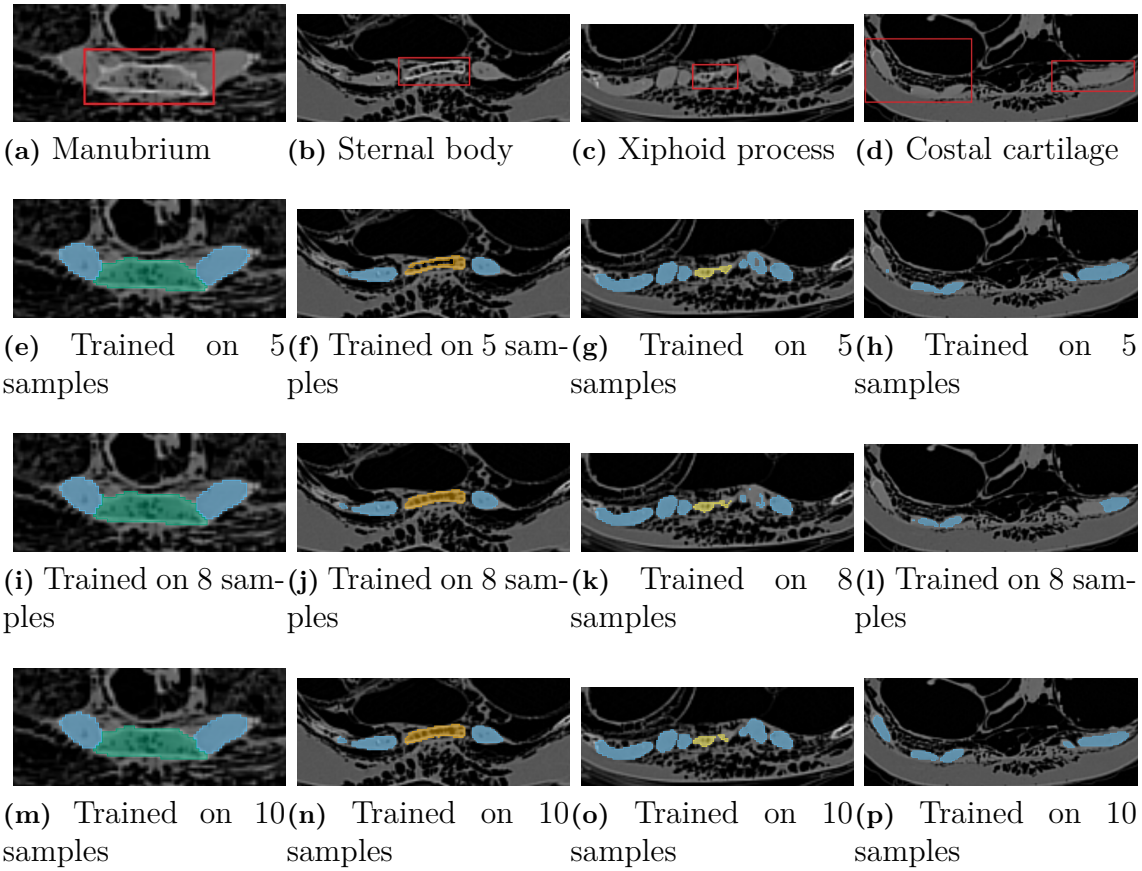
In contrast to the strong results on volume A, the model trained on the eight-sample dataset struggled with volume B. While the manubrium and sternal body were reasonably well segmented, Figure 4.5b, Figure 4.6i-j, the xiphoid process was only partially captured, and the costal cartilages were poorly segmented. This is particularly visible in Figure 4.6k-l, where large parts of the cartilage are missing.

Finally, the prediction from the model trained on the ten-sample dataset is shown in Figure 4.5c and Figure 4.6m-p. This model managed to segment all classes, with only minor inaccuracies in the xiphoid process, as seen in Figure 4.6o.



(a) Trained on 5 samples **(b)** Trained on 8 samples **(c)** Trained on 10 samples

Figure 4.5: Predicted segmentations of volume B by models trained on datasets with 5, 8, and 10 samples.



(a) Manubrium **(b)** Sternal body **(c)** Xiphoid process **(d)** Costal cartilage

(e) Trained on 5 samples **(f)** Trained on 5 samples **(g)** Trained on 5 samples **(h)** Trained on 5 samples

(i) Trained on 8 samples **(j)** Trained on 8 samples **(k)** Trained on 8 samples **(l)** Trained on 8 samples

(m) Trained on 10 samples **(n)** Trained on 10 samples **(o)** Trained on 10 samples **(p)** Trained on 10 samples

Figure 4.6: Axial view of different areas of the segmented volume, volume B.

The overall results are reflected in the predictions for volume A and volume B, shown in Figures 4.3, 4.4, 4.5, and 4.6. While the model encountered some challenges, the model trained on a dataset of ten samples ultimately produced high-quality segmentations. Still, a few volumes required more than minor corrections to be usable—three such cases are shown in Figure 4.7.

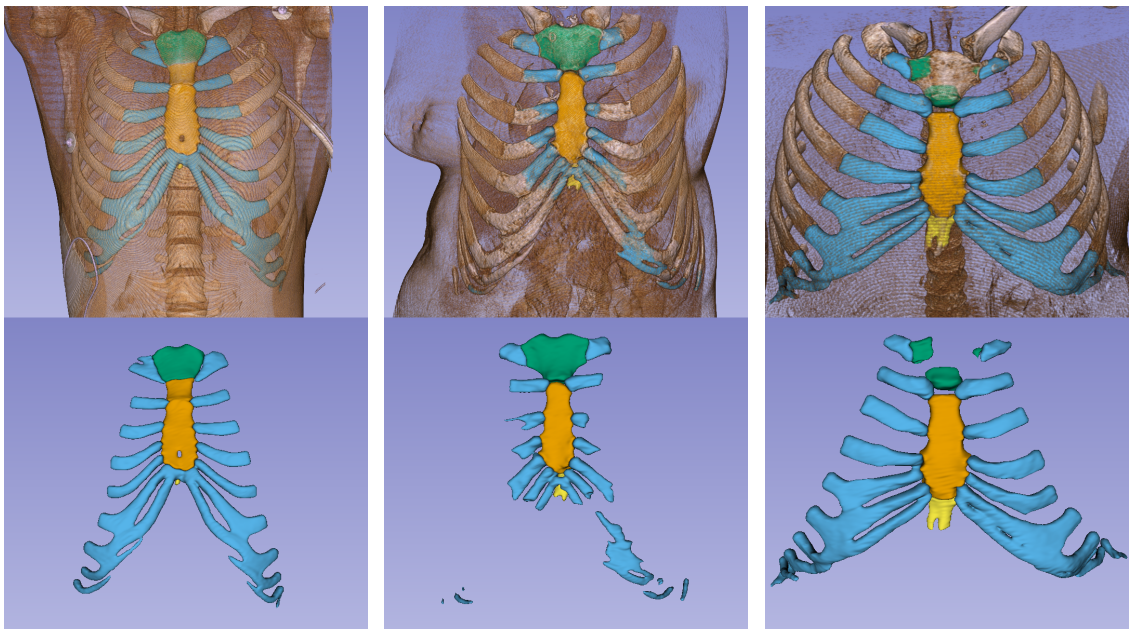
Despite generally strong performance, the model struggled with certain anatomical variations. Volume C (Figure 4.7a) shows an atypical sternum, where the manubrium is nearly as long as the sternal body. A split costal cartilage is also visible in the upper left corner. Volume D (Figure 4.7b) has close to fully calcified costal cartilages, and volume E (Figure 4.7c) features a ribcage that is broader than average.

(a) Volume C **(b)** Volume D **(c)** Volume E

Figure 4.7: Volumes where the model's prediction was less successful.

The predicted segmentations can be seen in Figure 4.8. The prediction of volume C has a too small manubrium since the second costal cartilages should be attached where the sternum switches from manubrium to sternal body, Figure 4.8a. It can also be seen that the upper left costal cartilage is almost split into two, which might have influenced the model to interpret it as cartilage one and two, and not only the uppermost one.

Figure 4.8b shows that the model had a hard time predicting the fully calcified cartilages in volume D. Further on the model struggled with the manubrium in volume E and failed to find the centerpiece although it managed to detect the edges connecting to the costal cartilage respective the sternal body, Figure 4.8c.



(a) Segmentation volume C **(b)** Segmentation volume D **(c)** Segmentation volume E

Figure 4.8: Outlying predictions, predicted by the model trained on 10 samples.

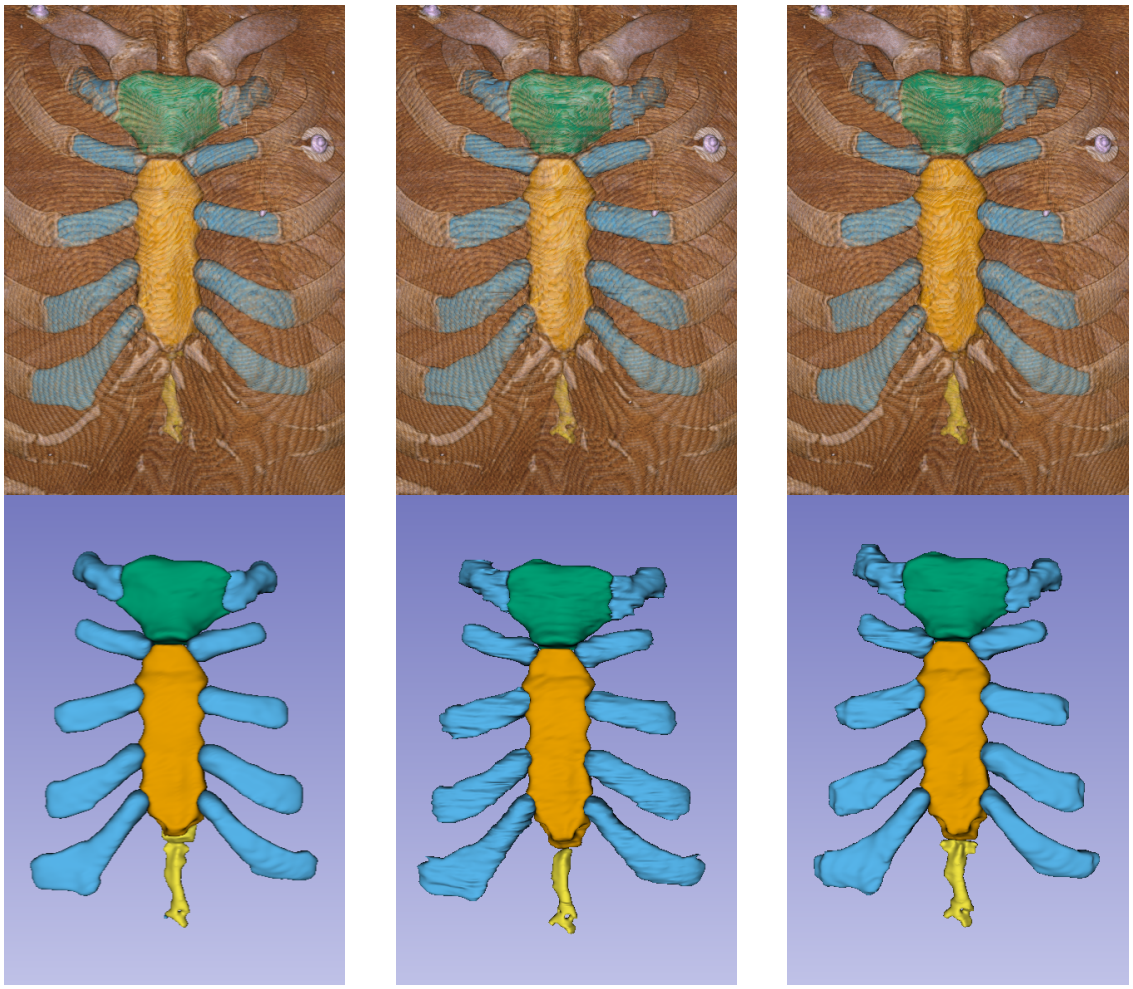
4.1.3 Model vs. manual segmentation: comparing to human performance

Volume F was chosen to be segmented manually as well as predicted with a model trained on 10 samples. The volume that was segmented can be seen in Figure 4.9. The manual segmentations and the model's prediction (after post-processing) are shown in Figure 4.10. The volume can be seen to have some calcification, mostly in the attachment of the costal cartilage to both the sternum and ribs. The sternum is otherwise considered normal.

Figure 4.9: Volume F, used for comparison between two human segmentations and a model prediction.

By comparing the model's prediction (Figure 4.10a) to the manual segmentations (Figures 4.10b–c) in 3D view, it is evident that the model produces a smoother and more consistent segmentation. The transitions between layers appear more continuous and anatomically natural. In the uppermost cartilage, which is almost fully calcified and nodular, the model successfully captures the nodular character while still generating a smoother result than either of the manual segmentations.

Figure 4.10 also illustrates that the model and the two individuals exhibit three different interpretations of the xiphoid process. Comparing Figures 4.10b and 4.10c reveals both similarities and differences. While the manubrium and sternal body appear relatively consistent at first glance, there are noticeable variations in the segmentation of the costal cartilages and xiphoid process. For example, differences can be seen in how the costal cartilage attaches to the sternal body and the ribs.



(a) Model prediction

(b) Individual 1, manual segmentation

(c) Individual 2, manual segmentation

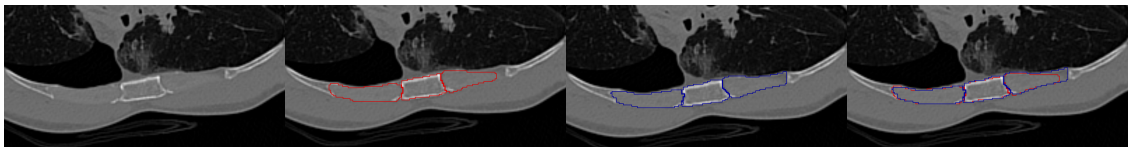
Figure 4.10: Comparison between the model's segmentation and the two manual segmentations.

The slice views in Figure 4.11 highlight how the manual segmentations differ from each other, even though they generally maintain a similar overall shape. In Figure 4.11a, the attachment of the cartilage to the rib on the right (reader's right) is segmented differently by the two individuals.

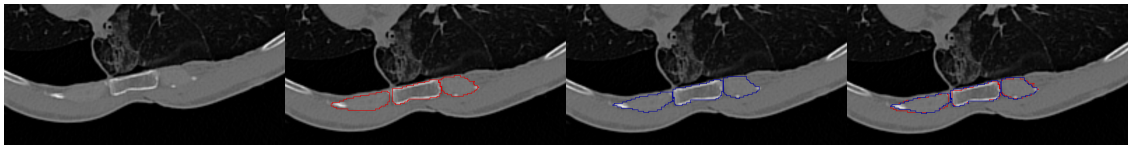
Despite these differences, there are also clear similarities. For example, Figure 4.11b shows good alignment between the segmentations. However, just as there are varying interpretations of the cartilage–rib attachment, Figure 4.11c illustrates differences in how the cartilage's attachment to the sternal body is segmented.

Table 4.3 supports these observations, showing that both individuals produced similar segmentations of the manubrium and sternal body, with some small variation in the xiphoid and greater differences in the costal cartilages. The model's segmentation aligns well with the human segmentations in the manubrium and sternal body but differs more notably in its interpretation of the xiphoid process and costal car-

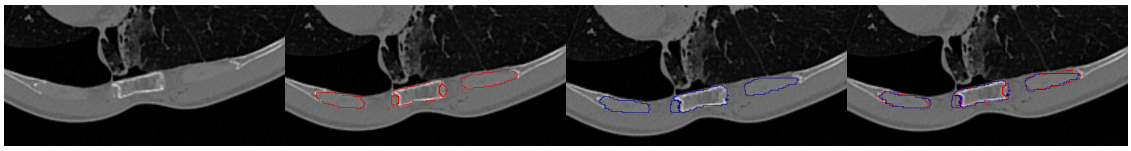
tilages. The per-class and total Dice coefficients in Table 4.3 show a similar pattern to those in Table 4.2.



(a) Slice showing different attachment of cartilage to rib



(b) Slice showing almost identical segmentation



(c) Slice showing different attachment of cartilage to sternal body

Figure 4.11: Comparison between two human segmentations, axial view from three different slices. First view only CT scan, second view individual 1, third view individual 2, fourth view both individuals.

Table 4.3: Dice coefficients per class and total, comparing segmentations by individual 2 and the model against those by individual 1.

Segmented by	Manubrium	Sternal body	Xiphoid process	Costal cartilage	Total
Individual 1	1	1	1	1	1
Individual 2	0.98903	0.99148	0.98577	0.94750	0.95441
Model	0.98228	0.96658	0.93928	0.93208	0.93915

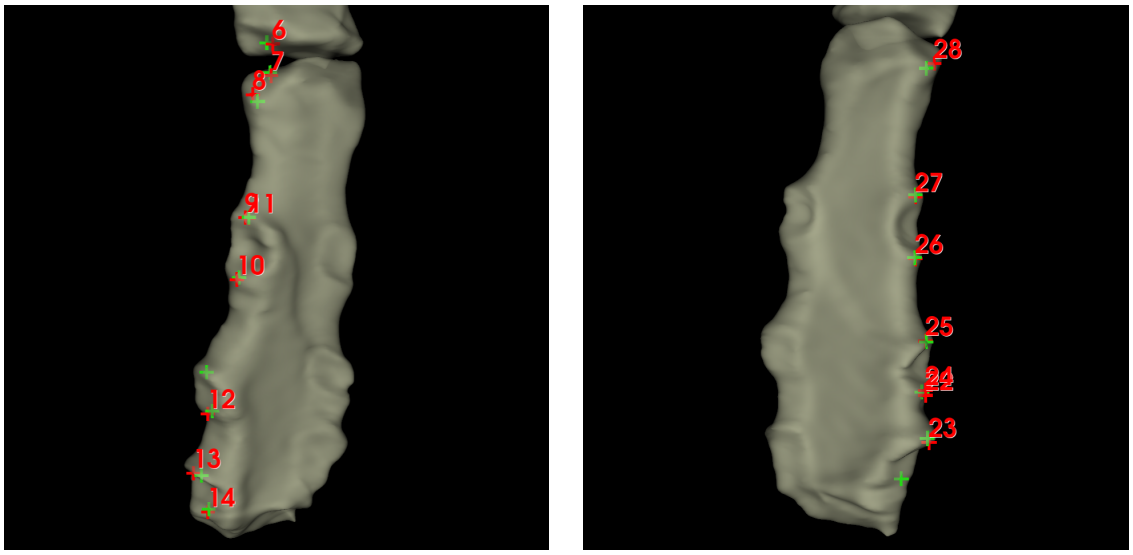
4.2 Landmarking results

The results indicate that deep learning can be successfully applied to the task of anatomical landmark localization. By leveraging both local image features and spatial relationships between landmarks, the proposed model was able to predict 3D landmark positions with high precision. This demonstrates that landmarking, like segmentation, is a task that can benefit from deep learning even when working with small datasets.

The results for landmark prediction are summarized in Table 4.4. With a dataset size of 20 samples, the model achieves a median inter-point error (IPE) of 1.68 mm, a mean of 2.82 mm, and a standard deviation (SD) of 4.31 mm. Approximately 60%

of the predicted landmarks fall within a 2 mm radius of the ground truth. Increasing the radius to 4 mm captures 87.5% of the predictions. However, around 5% of the predicted landmarks fall outside a 10 mm range and are considered misclassified.

These misclassifications occur when the spatial configuration component fails to resolve ambiguities between competing local predictions, leading to incorrect landmark assignments. Two such examples are illustrated in Figure 4.12, which shows misclassified landmarks in volume G. The landmarks are numbered in an anti-clockwise manner. In Figure 4.12a, both points 9 and 11 are predicted at the location of point 9. Similarly, in Figure 4.12b, points 22 and 24 are both assigned to the location of point 24. Since these outliers heavily impact the IPE metrics, the results are also reported with the misclassified points excluded in Table 4.5. When excluding these cases, the median IPE improves slightly to 1.61 mm, the mean decreases to 1.92 mm, and the standard deviation is reduced to 1.36 mm for a dataset size of 20 samples, indicating a consistent performance across correctly predicted landmarks.

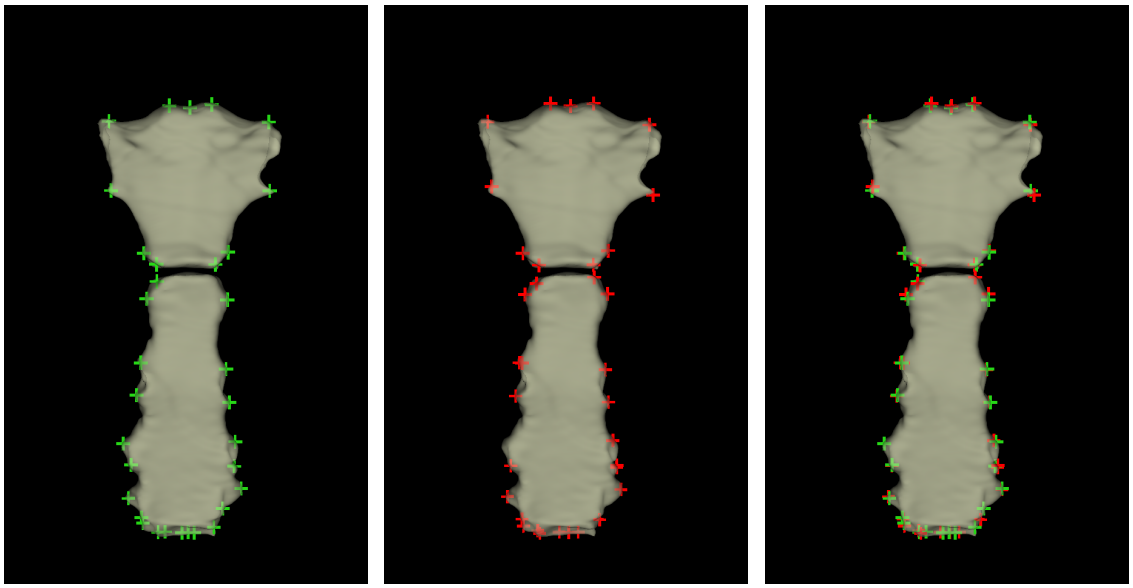


(a) Point 11 misclassified as point 9 **(b)** Point 22 misclassified as point 24

Figure 4.12: Landmarks misclassified by the model

The model showed greater prediction inaccuracy around the xiphisternal joint, the manubriosternal joint, and the lateral aspects of the manubrium, compared to the more consistent predictions along the sides of the sternal body and the superior edge of the manubrium, as illustrated in Figure 4.13c. However, while the landmarks along the lateral aspects of the sternal body are generally predicted with higher accuracy, they are also more prone to misclassification due to the repeated structure in that region, which can lead to confusion in landmark assignment.

While the 20-sample dataset showed promising results, performance notably degraded for smaller datasets. When reducing the dataset from 20 to 15 samples, there is a substantial increase in the number of misclassified landmarks. This can be seen in the last column of Table 4.4, where the proportion of landmarks predicted outside a 10 mm radius increases from 5.15% to 14.71%. However, when the



(a) Ground truth landmark points **(b)** Predicted landmark points from the model **(c)** Overlay of both ground truth and predicted points

Figure 4.13: Visualization of ground truth and predicted landmarks around the sternum for test volume G.

misclassified landmarks are excluded (Table 4.5), the model's performance becomes much more consistent across the 15 and 20 sample datasets. The median and mean IPE values remain close (1.46 mm vs. 1.61 mm median, and 1.97 mm vs. 1.92 mm mean), suggesting that the correctly predicted landmarks are estimated with comparable accuracy.

In contrast, further reducing the dataset to 10 samples leads to a higher rate of misclassifications, but perhaps more importantly, also a sharp decline in the precision of correctly classified landmarks. While the proportion of misclassifications only increases slightly (from 14.71% at 15 samples to 16.91% at 10 samples), the misclassification-free mean accuracy worsens notably — from 1.97 mm to 2.37 mm mean IPE, and outliers at $r = 2$ mm go from 32.76% to 53.10%.

Overall, these results indicate that the SCN-based landmarking approach substantially benefits from increasing the training dataset size. While 15 samples already reduce misclassifications and stabilize performance, 20 samples offer further improvements, both in the proportion of correctly classified landmarks and in the precision of those predictions. This highlights the importance of dataset size when applying deep learning models to tasks involving anatomical variability.

Table 4.4: Landmark prediction results across different dataset sizes on a test set of 4 samples.

Dataset size	**IPE** (in mm)		$\mathbf{O_r}$ (in %)		
	Median	Mean ± SD	r = 2mm	r = 4mm	r = 10mm
10	2.31	6.45 ± 10.20	61.03	27.21	16.91
15	1.75	5.23 ± 8.50	42.65	23.53	14.71
20	1.68	2.82 ± 4.31	38.24	12.50	5.15

Table 4.5: Landmark prediction results across different data set sizes, excluding 10 mm outliers. Calculated on a testset of 4 samples.

Dataset size	**IPE** (in mm)		$\mathbf{O_r}$ (in %)		
	Median	Mean ± SD	r = 2mm	r = 4mm	r = 10mm
10	2.09	2.37 ± 1.52	53.10	12.39	–
15	1.46	1.97 ± 1.53	32.76	10.34	–
20	1.61	1.92 ± 1.36	34.88	7.75	–

5

Discussion

This chapter discusses the results of the study and outlines suggestions for future work based on these findings.

5.1 Result analysis

This section analyzes the outcomes of applying deep learning to automatic segmentation and landmarking. It also discusses how the quality and consistency of manual annotations may have influenced the segmentation model's performance.

5.1.1 Using deep learning for automatic segmentation

The results presented in Section 4.1.2 suggest that deep learning can be used effectively for automatic segmentation of medical images with some exceptions. The Dice coefficients shown in Table 4.1 and Table 4.2, along with the segmentations in Figure 4.3 and Figure 4.5, indicate a high level of precision in the model's predictions. These results also show that overall segmentation accuracy tends to increase with larger datasets. However, this improvement is not consistent across all anatomical classes, as reflected in the Dice coefficients. It is important to note that a Dice score of 1 would mean that the prediction matches the manual segmentation exactly - but this is not necessarily ideal, as manual segmentations are unlikely to be perfect.

The model performed better on structures like the manubrium and sternal body, which are more regular in shape and structure, while it struggled with the xiphoid process and costal cartilage. These structures exhibit more anatomical variation, and in the case of the costal cartilage, differences in calcification and shape further complicate segmentation. The results indicate that these difficulties were reduced when the model was trained on larger datasets, likely because the model had more opportunities to learn from varied examples, and also because parts of the dataset were generated by the model rather than through manual segmentation. The results also reveal differences between the segmentations performed by two different individuals, particularly for the xiphoid process and the costal cartilages, highlighting the subjective nature of manual annotation

As mentioned in Section 3.3.4 *Expanding to final dataset*, the dataset was constructed through a selection process that excluded outliers. Additionally, the combinations available for forming datasets of five, eight, and ten samples were limited. The results

from volumes A and B (Figure 4.3, 4.5) suggest that both the specific features of a CT volume and its image quality influence the model's performance. They also emphasize that the choice of training and validation data plays a crucial role. A larger dataset would likely provide more variation and reduce the risk of excluding certain outlier types entirely. However, while increased variation gives the model more examples to learn from, it may also introduce noise, which can make learning and generalization more difficult. This trade-off was not explicitly explored in this study, but is worth investigating in future work.

Some of the less accurate predictions appear to be linked to features not represented in the dataset, for example, abnormal sternum shapes or fully calcified cartilages. However, in volume E, the model struggled to predict the manubrium (Figure 4.8a), even though this is typically one of the structures it segments well. This volume does not have clear outlier characteristics, apart from a relatively broad ribcage. Why this would affect the model's ability to segment the manubrium remains unclear. This underlines that simply increasing the dataset size may not be enough to solve all prediction challenges.

Besides dataset composition, the quality of the manual segmentations also influenced the results. Once the nnU-Net framework proved capable of automatic segmentation, efforts were made to refine the original manual annotations and to expand the costal cartilage segmentations from five to ten. These refinements addressed issues where the model had difficulty segmenting the cartilage–rib attachment areas and calcified regions. When trained on the improved dataset, the model showed enhanced performance in these challenging areas, regardless of dataset size.

5.1.2 The accuracy of manual segmentation

The accuracy of the manual segmentations used to train the model plays a key role in the results. These segmentations are created by humans, which means they can include small mistakes or differences in how details are interpreted. In this study, there were noticeable differences between the segmentations made by two individuals, even though both were treated as the correct reference. The individuals had different interpretations of the xiphoid process and the costal cartilages, suggesting that it is a particularly challenging structure to segment and that the manual annotations used for training may have lacked consistency.

One way to improve the dataset could be to have several people review or adjust the segmentations to make them more consistent. However, it is a balance between how much effort is spent on creating the dataset and how much that extra accuracy actually improves the model's performance.

5.1.3 Using deep learning for automatic landmarking

The results indicate that deep learning can be successfully applied to the task of anatomical landmark localization. By leveraging both local image features and spatial relationships between landmarks, the proposed model was able to predict 3D landmark positions with high precision. This demonstrates that landmarking,

like segmentation, is a task that can benefit from deep learning even when working with a small dataset.

With a dataset of 20 samples, the model achieved strong overall performance, as shown in Table 4.4. It reached an IPE of 1.68 mm, with 87.5% of predictions falling within a 4 mm radius of the ground truth. This level of accuracy highlights the model's ability to integrate local features with spatial configuration effectively. However, the relatively high standard deviation (4.31 mm) and mean IPE (2.82 mm) suggest that a small number of outlier predictions substantially degrade the overall performance.

Closer analysis revealed that approximately 5% of the landmarks were misclassified, which in this case meant that two landmarks were assigned to the same location. These cases occur when the spatial configuration module fails to disambiguate between possible local predictions, particularly in anatomically repetitive areas. When these misclassified predictions were excluded from evaluation, the mean IPE dropped to 1.92 mm, and the standard deviation was reduced to 1.36 mm, indicating that most predictions are accurate and consistent. Notably, manually correcting 0–2 misclassified landmarks per volume is relatively quick, supporting that a semi-automatic approach remains a viable solution. This workflow could still substantially reduce the overall annotation burden while maintaining high accuracy.

As with the segmentation predictions, the size of the dataset has a large impact on model performance. Reducing the training dataset from 20 to 15 samples resulted in a near threefold increase in misclassifications (from 5.15% to 14.71%), though the accuracy of correctly predicted landmarks remained comparable. This suggests that with a smaller dataset, the model remains accurate when confident but is more prone to making misclassification errors. When reducing the dataset further to 10 samples, both the misclassification rate and the error of correctly predicted landmarks worsened. Even after excluding misclassified points, the mean IPE rose from 1.97 mm (at 15 samples) to 2.37 mm, and the proportion of points falling outside a 2 mm radius jumped from 32.76% to 53.10%. These findings imply that the model begins to overfit with very limited data, losing generalizability to unseen anatomical variation.

These results emphasize the importance of dataset size in training robust models for anatomical landmark detection. While 15 samples represent a significant improvement over 10, a dataset of at least 20 samples substantially improves both the number and precision of correct landmark predictions. The results indicate that even modest increases in data size can yield disproportionately large performance gains. This supports an iterative strategy in which model predictions are manually corrected and added into the training set, progressively improving the model as the dataset grows.

Anatomical variability and segmentation accuracy also played a part in the model's performance. The model was less accurate around the xiphisternal and manubriosternal joints, as well as the lateral aspects of the manubrium. These regions exhibit greater variability between subjects, making them harder to localize precisely. In contrast, predictions along the lateral aspects of the sternal body and the superior

edge of the manubrium were more consistent. However, the lateral aspects of the sternal body, despite being predicted with higher accuracy on average, were more susceptible to misclassification. This is likely due to the repetitive structure along the length of the sternal body, which increases the difficulty of assigning the correct landmark identity when similar local features are present nearby.

One limitation of the current implementation lies in the heatmap-based pre-processing approach. Instead of applying spatial augmentations directly to the landmark coordinates, the augmentations during training were applied to entire heatmap volumes. This approach significantly increased computational cost, as full 3D volumes had to be transformed rather than a small set of points. Augmenting heatmaps rather than coordinates also introduced the risk of discretization artifacts due to interpolation and voxel misalignment, especially when landmarks were transformed to non-integer locations that did not align cleanly with the voxel grid. These interpolation effects can blur or shift the peak of the heatmap, thereby degrading the precision of the training signal. A more precise and efficient alternative would be to transform the continuous landmark coordinates using the same spatial parameters as the image augmentations, and then generate the heatmaps afterwards. This would reduce both computational overhead and the likelihood of introducing discretization-related artifacts. Upsampling the image to a higher resolution is another potential mitigation strategy, but it comes with increased VRAM and training time requirements.

Overall, the results show that the proposed deep learning approach can deliver high accuracy even with limited training data. With minimal manual correction, it can be used as part of a semi-automatic pipeline that balances efficiency and precision, making it a practical solution for anatomical landmarking in resource-constrained settings. Furthermore, insights into anatomical ambiguity and the impact of preprocessing choices offer actionable guidance for future improvements.

5.2 Future work

This study focuses on how dataset size influences deep learning models' abilities to perform segmentation and landmarking on medical images. Future work could explore how the degree of variation within the dataset affects performance. It would be valuable to investigate whether there is an optimal level of variation - one that gives the model enough variation to learn from, without making it harder to recognize useful patterns. Gaining a better understanding of how variation affects learning could lead to improved segmentation results, as a well-balanced dataset may make the model more robust to outliers and uncommon cases.

Another area for improvement involves the consistency of segmentation methods used in the datasets. In this study, the five-sample dataset was manually segmented, while the ten-sample dataset included five semi-automatically produced segmentations. To make the results more trustworthy and comparable, future experiments should use the same segmentation method across all dataset sizes. One possible approach would be to train the model on five manual segmentations and then use it to generate additional ones, creating datasets of five, eight, and ten samples with

consistent methodology.

Currently, predicted segmentations need to be manually checked to catch outliers. Automating this would help reduce manual effort, but it's not straightforward since there's no ground truth to compare against. One idea is to define acceptable size ranges for each class based on anatomical knowledge, such as limits on width, height, and depth. This approach relies on static thresholds and may not fully capture natural variation.

It might also be possible to compare predictions to one or more anatomical templates of the sternum and costal cartilages. Using a set of representative topologies, rather than a single average-shaped template, could better account for normal anatomical variation and improve the robustness of outlier detection. However, even with multiple templates, some outliers might still be missed due to extreme anatomical differences. Therefore, to reliably detect all outliers, a more flexible or combined approach might prove necessary.

Another potential direction for future work is to further generalize the approach to make it more user-friendly, ideally resulting in a modular, plug-and-play solution that can be easily adapted to different anatomical regions without the need for extensive reconfiguration. Improving the usability of the method would make it more accessible to users with varying levels of technical expertise and broaden its potential applications. Additionally, the landmarking process could be expanded to cover a larger area, including the costal cartilages.

The method for automatic landmarking could also be extended to incorporate non-anatomical landmarks, known as pseudo landmarks. These are evenly spaced points placed between anatomical landmarks to better capture the overall geometry of a structure. Such landmarks can be generated using a closed-form solution based on the predicted anatomical points.

Improvements to the landmarking pre-processing pipeline are necessary to enable generalization and scalability. Specifically, the current method of updating landmark coordinates by augmenting full 3D Gaussian heatmaps introduces both interpolation artifacts and considerable computational overhead. This becomes a bottleneck when applying more complex augmentations, such as elastic deformations, or when increasing the spatial dimensions of the input data. Replacing this step with a more efficient approach, such as applying the spatial transformations directly to the physical landmark coordinates, would drastically reduce training time, support a broader range of augmentations, and make it feasible to extend the method to larger anatomical regions.

6

Conclusion

This thesis aimed to investigate the feasibility and performance of deep learning models for segmentation and landmarking of the sternum and costal cartilage, particularly under constraints of limited annotated data. The study addressed the following research questions:

- How can a deep learning model be adapted for sternum and costal cartilage segmentation?
- What level of accuracy can be achieved with a limited amount of labeled data?
- How does the amount of labeled data affect the resulting segmentation?
- How can the approaches and insights from segmentation be extended to automatic landmarking of the sternum?

When evaluating the segmentations produced by nnU-Net trained on datasets consisting of five, eight, and ten samples respectively, it can be concluded that the model achieves a high level of accuracy, even when trained on as few as ten samples. Despite the limited dataset size, the model delivers consistent and reliable results. Notably, the model trained on five or eight samples tended to miss certain anatomical structures that were correctly segmented when trained on ten samples. Systematic errors decreased with larger datasets, and the model demonstrated a greater ability to capture the overall structure.

A comparison between nnU-Net's segmentations and manual segmentations shows that the model is not only more consistent but also more efficient. This includes the time required for dataset preparation, inference, post-processing, and manual correction of inaccuracies. Since the model was mainly trained on a set of ten examples, it does not generalize to all possible anatomical irregularities, making a semi-automatic approach necessary to ensure accuracy. In this context, "semi-automatic" refers to using model-generated segmentations followed by manual review and correction, as well as optional retraining of the model with a larger dataset.

While manual segmentation may still outperform automated methods in outlier cases, the semi-automatic approach often provides a better trade-off between accuracy and efficiency. Future work should aim to improve the model's robustness to rare anatomical variations to enable fully automatic segmentation (see Section 5.2).

Moreover, insights from the segmentation task informed the overall pipeline design and motivated the choice of SCN as a suitable model architecture for the landmark-

ing task. Models were trained on datasets containing 10, 15, and 20 samples. With 20 training samples, the model achieved a median inter-point error (IPE) of 1.68 mm and a mean IPE of 2.82 mm, with over 87% of predictions falling within a 4 mm radius of the ground truth. While most predictions were accurate, around 5% of points were misclassified. When these points were excluded, the mean IPE improved to 1.92 mm with reduced standard deviation, confirming the model's precision.

As with segmentation, performance degraded as the number of training samples decreased. Reductions in dataset size led to higher localization errors and increased misclassification rates, highlighting the sensitivity of deep learning models to training data volume in both tasks.

Because both models require only minimal manual correction to achieve high accuracy, a semi-automatic workflow is well supported for both tasks. Model predictions can be reviewed, corrected as needed, and either used directly or incorporated into an expanded training set to iteratively improve model performance.

The findings support the conclusion that deep learning methods, using nnU-Net and SCN, are well-suited for medical image analysis tasks like segmentation and landmarking, even when annotated data is scarce. In both cases, a semi-automatic approach, where predictions are reviewed and corrected, enables efficient and accurate results. Moreover, both models show a consistent trend: larger training datasets improve model robustness and reduce the need for manual correction.

Future work should aim to improve model robustness to rare anatomical variations and ensure consistency in dataset preparation to enable fully automatic segmentation without manual review. For landmarking, improving the pre-processing pipeline could reduce computational overhead and allow the method to scale to larger anatomical regions.

Bibliography

[1] Trafikverket. "This is vision zero." (2020), [Online]. Available: https://bransch.trafikverket.se/en/startpage/operations/Operations-road/vision-zero-academy/This-is-Vision-Zero/ (visited on 2025-01-30).

[2] S. Koppel, J. Charlton, B. Fildes, and M. Fitzharris, "How important is vehicle safety in the new vehicle purchase process?" *Accident Analysis & Prevention*, vol. 40, no. 3, pp. 994–1004, 2008, ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2007.11.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457507002023.

[3] R. Ren, H. Li, T. Han, *et al.*, "Vehicle crash simulations for safety: Introduction of connected and automated vehicles on the roadways," *Accident Analysis & Prevention*, vol. 186, p. 107 021, 2023, ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2023.107021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001457523000684.

[4] J. Iraeus, K. Brolin, and B. Pipkorn, "Generic finite element models of human ribs, developed and validated for stiffness and strain prediction – to be used in rib fracture risk evaluation for the human population in vehicle crashes," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 106, p. 103 742, 2020, ISSN: 1751-6161. DOI: https://doi.org/10.1016/j.jmbbm.2020.103742. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1751616120302964.

[5] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: A review," *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.

[6] A. A. Weaver, S. L. Schoell, C. M. Nguyen, S. K. Lynch, and J. D. Stitzel, "Morphometric analysis of variation in the sternum with sex and age," *Journal of morphology*, vol. 275, no. 11, pp. 1284–1299, 2014.

[7] H. J. H. Edgar, S. Daneshvari Berry, E. Moes, N. L. Adolphi, P. Bridges, and K. B. Nolte, *New mexico decedent image database*, 2020. DOI: 10.25827/5s8c-n515.

[8] Z. Zhou and S. Liu, *Machine Learning*. Springer Nature Singapore, 2021, ISBN: 9789811519673. [Online]. Available: https://books.google.se/books?id=ctM-EAAAQBAJ.

[9] K. P. Murphy, *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013, ISBN: 9780262018029 0262018020. [Online]. Available: https://www.amazon.com/Machine-Learning-Probabilistic-

Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2.

[10] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021-09. DOI: 10.1007/s12525-021-00475-2. [Online]. Available: https://doi.org/10.1007/s12525-021-00475-2.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[13] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[14] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024-03, ISSN: 1573-7462. DOI: 10.1007/s10462-024-10721-6. [Online]. Available: https://doi.org/10.1007/s10462-024-10721-6.

[15] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024-03.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241.

[17] F. Isensee, J. Petersen, A. Klein, *et al.*, "Nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[18] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, "Continuous dice coefficient: A method for evaluating probabilistic segmentations," *arXiv preprint arXiv:1906.11031*, 2019.

[19] A. Galdran, G. Carneiro, and M. A. G. Ballester, "On the optimal combination of cross-entropy and soft dice losses for lesion segmentation with out-of-distribution robustness," in *Diabetic Foot Ulcers Grand Challenge*, Springer, 2022, pp. 40–51.

[20] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., Cham: Springer International Publishing, 2016, pp. 230–238.

[21] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based cnns for landmark localization," *Medical Image Analysis*, vol. 54, pp. 207–219, 2019, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2019.03.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841518305784.

[22] T. DenOtter and J. Schubert, *Hounsfield unit*, In: StatPearls [Internet], [Updated 2023 Mar 6], Treasure Island (FL), 2023-03. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK547721/.

[23] DICOM Standards Committee, *About DICOM: Digital Imaging and Communications in Medicine*, Accessed: February 4, 2025, 2025. [Online]. Available: https://www.dicomstandard.org/about.

[24] M. Mustra, K. Delac, and M. Grgic, "Overview of the dicom standard," in *2008 50th international symposium ELMAR*, IEEE, vol. 1, 2008, pp. 39–44.

[25] National Electrical Manufacturers Association (NEMA), *DICOM Standard - Section C.7.6.2.1.1: Image Plane Module*, Online, Accessed: 2025-04-02, 2013. [Online]. Available: https://dicom.nema.org/dicom/2013/output/chtml/part03/sect_C.7.html#sect_C.7.6.2.1.1.

[26] G. Kindlmann, *Nrrd: Nearly raw raster data file format*, Online, Accessed: 2025-04-02, 2025. [Online]. Available: https://teem.sourceforge.net/nrrd/format.html.

[27] G. Kindlmann, *Nrrd file format specification*, https://teem.sourceforge.net/nrrd/format.html, Accessed: 2025-04-02, 2025.

[28] Y. Shafranovich, "Common format and mime type for comma-separated values (csv) files," Tech. Rep., 2005.

[29] D. R. Haase and H. S. Shaikh, "Anatomy of the ribs, sternum, and costal margin," *Journal of Orthopaedic Trauma*, vol. 38, no. 12S, S1–S6, 2024.

[30] J. L. Forman and R. W. Kent, "The effect of calcification on the structural mechanics of the costal cartilage," *Computer methods in biomechanics and biomedical engineering*, vol. 17, no. 2, pp. 94–107, 2014.

[31] S. Katina, K. McNeil, A. Ayoub, *et al.*, "The definitions of three-dimensional landmarks on the human face: An interdisciplinary view," *Journal of Anatomy*, vol. 228, no. 3, pp. 355–365, 2016. DOI: https://doi.org/10.1111/joa.12407. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joa.12407. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/joa.12407.

[32] 3D Slicer Community, *3D Slicer: Open-Source Platform for Medical Image Computing*, Accessed: February 4, 2025, 2024. [Online]. Available: https://www.slicer.org.

[33] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, *et al.*, "3d slicer as an image computing platform for the quantitative imaging network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, 2012-11. DOI: 10.1016/j.mri.2012.05.001.

[34] C. Pinter, A. Lasso, and G. Fichtinger, "Polymorph segmentation representation for medical image computing," *Computer Methods and Programs in Biomedicine*, vol. 171, pp. 19–26, 2019, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2019.02.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260718313038.

[35] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020-09. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[36] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of SimpleITK," en, *Front Neuroinform*, vol. 7, p. 45, 2013-12.

[37] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[38] F. Pérez-García, R. Sparks, and S. Ourselin, "TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, p. 106 236, 2021, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106236. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260721003102.

[39] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. DOI: 10.1038/s41592-020-01008-z.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014, Cited by: 33793. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84904163933&partnerID=40&md5=b865fd654b3befc5d829dbe5d42b80c3.

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY