



FACULTY OF
COMPUTER SCIENCE

# CNN-BASED SPINAL CURVATURE RECONSTRUCTION/EXTRACTION IN MRI FOR STATISTICAL ANALYSIS IN POPULATION STUDIES

PHILIPP ERNST

Master's Thesis
Department of Simulation and Graphics
Faculty of Computer Science
Otto von Guericke University Magdeburg

Supervisors:
Prof. Dr.-Ing. Klaus Tönnies
Dr. Marko Rak

October 2018

## STATEMENT OF AUTHORSHIP

I hereby declare that I am the sole author of this master thesis and that I have used only the sources listed in the bibliography and identified as references. I further declare that I have not submitted this thesis at any other institution to obtain a degree.

*Magdeburg, October 2018*

Philipp Ernst

## ABSTRACT

In this thesis, four strategies for a fast and accurate extraction of the spinal curve from MR records using CNNs are presented and evaluated. Furthermore, the extracted spines from a dataset of more than 3000 records (SHIP dataset) are tested statistically on how certain biological features like age or weight influence the curvature. Additionally, ranges of Cobb angles for healthy spines are determined and a visual tool for a fast classification of the spine is proposed.

It turns out that the best strategy results in a maximal mean error of only 3.3 mm and the healthy Cobb angles range from 40° to 50° and -7° to 9° in the sagittal and coronal view, respectively. The statistical tests interestingly showed, among other insights, that it is rather unlikely/unnatural to have a straight but rather slightly bent spine in the coronal view, i.e. mild scoliosis.

# CONTENTS

# ACRONYMS

| | |
|---|---|
| KS | Kolmogorov-Smirnov |
| MWU | Mann-Whitney-U |
| CNN | Convolutional Neural Network |
| CPR | Curved Planar Reformation |
| MR | Magnetic Resonance |
| MRI | Magnetic Resonance Image |

# 1 INTRODUCTION

In this chapter, a general introduction to the subject of this thesis will be given. Therefore, the need of this work is motivated in the first section of this chapter. Thereafter, the goal for the algorithm that will be proposed in a later chapter is delimited.

## 1.1 MOTIVATION

The vertebral column of human beings (and vertebrates in general) is the part of the body that primarily supports the statics of the whole body. It consists of 33 osseous vertebral bodies and intervertebral discs that lie between most of these vertebrae. The whole spine can be divided into the cervical, thoracic and lumbar region followed by the sacrum and the coccyx from top to bottom. The spine of a healthy subject should, in a coronal view, be straight. In a sagittal view, a healthy spine is not straight but rather curved to better stabilize the whole body. This curvature is not arbitrary but dorsal concave (called 'lordotic curve') in the cervix, dorsal convex (called 'kyphotic curve') in the thorax and again lordotic in the lumbar region. The curvature in the sacral and coccyx region can be described as kyphotic[6].

For thousands of years, spinal deformities have impaired the life of people living with such a disease[48]. Due to that fact, it is understandable that doctors want to treat these malformations as fast and as good as possible. Once people have reached adulthood, the spine will hardly change any more which means that it is unlikely to develop a spinal deformity in the years of adulthood but also that a malformation untreated during adolescence can hardly be cured as an adult[5].

This means that it is important for doctors to know when to start a treatment, i.e. the borders between classifying a spine as healthy and unhealthy must be known.

There are several different types of spinal deformity that can occur. One of these, and probably the most common malformation[22], is called scoliosis. The spine of people diagnosed with scoliosis most often has a lateral shape like the letters 'C' or 'S' (possibly mirrored). If scoliosis occurs during adolescence and without an obvious reason it is called *adolescent idiopathic scoliosis (AIS)*. With 2 % - 3 % of children between 10 and 18 years affected by this malformation (especially girls)[3], AIS is a very prevalent. Once it reaches a certain degree, the patients can suffer from pain in the back, the risk of developing shortness of breath is increased and the self-esteem

can lower drastically[2]. Hence, a treatment with a brace or, in severe cases, surgery with spinal implants that hold the spine in a straight position is necessary.

Another type of spinal malformation is called *kyphosis*. In contrast to scoliosis, the curve of the spine is abnormally kyphotically convex, i.e. sagittally displaced outwards, which can occur in cervical, thoracic and sacral regions[41]. A mild kyphosis can only be diagnosed very difficultly since a healthy spine always has a kyphotic curvature which can also be changed very easily by not standing, sitting or lying completely straight[15]. Like with scoliosis, a high degree can cause pain, breathing difficulties and even shortened life spans. It is possible to cure patients with kyphosis by using a body brace, physical therapy or surgery which is very similar to the treatment of scoliosis.

The last spinal deformity presented in this chapter is termed *lordosis*. Such like kyphosis, this malformation is a sagittal displacement as well, but in this case, it describes the abnormally lordotic concave curvature of the spine and can occur in thoracic and sacral regions. Like scoliosis and kyphosis, lordosis can lead to back pain as well. Another symptom is spinal disc herniation which means that intervertebral discs have been damaged. Different to scoliosis and kyphosis, cases of lordosis usually do not occur due to inherent physical defects but rather because of habitual poor posture[28]. This is why it is usually treated with braces or stretches and not by surgery.

To both terms kyphosis and lordosis, the descriptions *hyper-* or *hypo-* can be attached to explain a rather strong convexity/concavity or to be rather straight, respectively. This is not necessary for scoliosis since a lateral curvature of the spine is not natural and thus would always be described as unhealthy.

Since these malformations are very common and need to be cured, there need to be methods to measure and classify the strength of these presented curvatures. The most common process is the classification after the determination of the *Cobb angle*[9], which was introduced to measure scoliosis but can also be applied to lordosis and kyphosis. In case of scoliosis, one needs a coronal view of the spine where all vertebral bodies are observable. The first step is to determine the neutral vertebrae. These are the vertebral bodies at the inflection points of the lateral curvature. The next step is to find the tangent at the superior endplate of the upper neutral vertebra and the tangent at the inferior endplate of the lower neutral vertebra, respectively. The angle between these two tangents is defined as the Cobb angle. If this value is 30° or less, the scoliosis is classified as mild and up to 45° it is moderate. Everything above is defined as severe scoliosis[4].

Similarly, the Cobb angle can also be determined in sagittal views of the spine to classify cases of lordosis or kyphosis. As stated before, a healthy spine is not straight when being viewed sagittally and

so the classification ranges of Cobb angles for scoliosis cannot be applied here. Rather, a Cobb angle between 20° and 45° from the first to the twelfth vertebra seems to be normal regarding kyphosis although the values change with age and gender[15]. This means, patients with Cobb angles greater than 45° are classified as hyperkyphotic and smaller than 20° as hypokyphotic.

It is even harder to diagnose lordosis or to specify a range of normal lordotic angles since it is not clear where to measure the angles. Some studies mention ranges from 20° to 60° or 40° to 60° and even up to 83.5°[43].

In cases of scoliosis, the Cobb angle alone often is not enough to precisely classify the disease since it does not tell anything about the direction or the general shape. This is why other methods have evolved that use the Cobb angle but also other information.

The first method was proposed by King et al.[21] and provides five different types that scoliosis can be classified as. Unfortunately, it is not possible to quantify the degree of these types with King's method. However, Lenke et al. [24] presented another method for not only describing the overall shape of the scoliosis but also providing some modifiers based on the Cobb angles. Yet another classification system was introduced by Rigo et al. [37] and classifies the spine in five different types, such like King et al. did but one needs to consider clinical data, e.g. 'pelvis translated to the concave thoracic side', as well as radiological criteria, e.g. 'positive L5-4 counter-tilting'.

All in all, there exist some methods to coarsely determine spinal malformations but they are neither tailored to the patient, i.e. except for Rigo et al., no method uses additional information next to the shape of the spine (like age, sex, weight, ...), nor allow to precisely classify the spine, e.g. by defining exact decision boundaries for healthy and unhealthy. Of course, this needs a big dataset and a fast possibility to analyze the data, which was not possible computationally until now.

## 1.2 GOAL

As one can see from the previous section, there are several different methods for classifying the shape of the spine, which overlap in some aspects but are rather different in other aspects (even for describing the same shape).

Furthermore, the ranges of Cobb angles, for classifying scoliosis as well as for classifying kyphosis and lordosis, seem to be chosen arbitrarily and only coarsely reflect the certain degrees of severity of a possible malformation. Finding more precise ranges and allowing for a statistically more accurate classification requires the analysis of a big representative dataset which is almost impossible with the described methods since it would be extremely time consuming.

On the other hand, the analysis of such a data set would not only permit to determine the ranges for the entirety of subjects but also to filter for certain properties like age or weight, if necessary. At last, this would allow for a uniform classification system.

For these reasons, this thesis aims to:

1. extract the curve of the spine from medical images in a fast and yet sufficiently accurate way (i.e. being able to extract hundreds or thousands of spinal curves in a short time),
2. find a method to map the extracted spinal curves to a common system to allow a facilitated comparison,
3. carry out statistical tests on the spines of a big dataset to find differences between groups of subjects with certain properties (e.g. young vs. old subjects),
4. determine ranges of Cobb angles to statistically and meaningfully distinguish between healthy and abnormal spine curvatures,
5. provide a visualization of the spine curvature to identify the region of the spine where critical values of Cobb angles are exceeded at a glance.

## 1.3 STRUCTURE OF THIS WORK

This chapter gave a coarse insight into the topic and the goals of this thesis.

Chapter 2 presents currently available methods for extracting spines from medical images and to classify the different diseases of the spine automatically.

In Chapter 3, the method of this thesis for extracting the centerline of the spine will be explained. Therefore, four strategies are presented that are similar but can lead to different outcomes nevertheless.

For this reason, Chapter 4 contains tests of all of the presented methods and their substeps on a real dataset. Based on the results, the best, i.e. most accurate and visually pleasing, method is chosen.

In Chapter 5, hypotheses on the spinal curves are stated and statistical tests are carried out on the extracted spines by the best method chosen in the previous chapter and on a big dataset of more than 3000 subjects to verify the validity of the hypotheses. Furthermore, a method for distinguishing between healthy and abnormal spine curvatures is proposed.

Finally, Chapter 6 summarizes the results of the previous chapters and suggests some future work that would be interesting for both doctors and computer scientists.

# 2

# STATE OF THE ART

Up until now, there have been made a lot of attempts to tackle problems in medical image analysis to classify many different diseases referring to the spine. This chapter gives an overview over certain methods for solving these problems.

## 2.1 SPINE DETECTION AND EXTRACTION

For computer-aided diagnosis, it is necessary to examine the spine in some cases. To do so, one needs to detect, localize or segment the vertebral bodies or the spine as a whole.

One attempt was made by Kelm et al.[20]. They propose an algorithm that can work on both CT or Magnetic Resonance (MR) records. It is a combination of Marginal Space Learning and a generative anatomical network, and consists of four steps: rough spine localization, disk selection (by iterated Marginal Space Learning), disk selection & labeling and structure segmentation. Using these steps, they cannot only estimate the location, orientation and size of each vertebral body but also carry out a segmentation using this information. The method obtains a sensitivity of $\geq$98 % with a processing time of 11.5 s on a dual-core 2.2 GHz Laptop.

Zukić et al.[55] presented a method for the detection and segmentation of vertebral bodies with minimal user intervention in MR records. The first step of this method is the detection of vertebral centers using the Viola-Jones algorithm (which needs to be trained). To eliminate outliers in the set of these initial detections, a third-order spline is approximated using the center positions and detections farthest away from this spine are removed. After this step, it is possible for the user to modify the set of center points. The next step is the segmentation of the vertebrae. Therefore, a probability map for the boundaries of the bodies is calculated using several different features like the results of an edge detection, a distance transform or a morphological closing. For every vertebral body, an iterative inflation algorithm using balloon forces is applied to finally segment them. Afterwards, these segmentations and/or the centers can be used to diagnose diseases like scoliosis, spondylolisthesis, or vertebral fracture. With a detection rate of $>$92 % and a Dice coefficient of $>$79 %, this method is capable of segmenting the spine in around 70 s for a whole dataset.

Chen et al.[7] take another approach of localizing and identifying vertebrae. Their method is a combination of a coarse detection us-

Figure 2.1: Deep Supervision

ing random forests for finding vertebra candidates and then feeding these into a Convolutional Neural Network (CNN) to predict the centroids of the vertebral bodies. As a modification to the standard CNN, they introduce the J-CNN which also takes pairwise conditional dependencies of neighboring vertebrae into account. Since these predictions could have offsets due to a low resolution of the vertical axis of their data, Chen et al. trained a shape regression model after the J-CNN to refine the coordinates of the predicted vertebrae. The method was trained and tested on 302 annotated CT scans and improved the standard CNN identification rate by around 10 %, achieving 84 % for the whole spine and even 92 % for the cervical region.

Another interesting attempt was made by Yang et al.[51] for vertebra labeling using encoder-decoder networks. This method also consists of several steps. The first one initializes the vertebra locations using a so-called deep image-to-image network similar to a U-Net[40] but extended by deep supervision (see Fig. 2.1), i.e. not only the final output but also the outputs of the intermediate layers are supervised and trained. The output of this network is a probability map of vertebra centers which are then iteratively evolved using message passing themes and the mutual relation of the centroids. The last step is a refining step like in [7] but using sparsity regularization. To their knowledge, they were the first ones who tested on a dataset of more than 1000 samples and achieved an identification rate of 90 % for the whole spine.

A method that is also based on neural networks was proposed by Wu et al.[50]. The algorithm works on spinal X-ray records and outputs four landmarks for each vertebral body depicted in the record. In its essence, this method is a regular CNN with a fully connected layer before the output layer but in addition also includes the newly presented BoostLayer after the convolutional layers. This new layer enhances the feature space by removing outliers and thus minimizes the intra-class variance. Wu et al. tested this method on 481 spinal X-ray images with manually extracted landmarks for training and test-

ing. These landmarks have been normalized to be in the range $[0, 1]$ with $(0.5, 0.5)$ being the center of the image. The method results in a mean squared error of only 0.0046 pixels compared to 0.018 pixels using the regular CNN without the BoostLayer.

Yao et al.[52] presented another interesting method to extract and partition the spinal column from CT records. It consists of five steps: initial spine segmentation (a thresholding at 200 HU followed by a connected component analysis retaining only the largest component), spinal canal extraction (by applying the watershed algorithm on every axial slice followed by a merging step and a directed graph search along the transverse axis), fitting of a four-part vertebral model (on every 2D slice by a maximum model-to-image match), Curved Planar Reformation (CPR) (to provide clear views of the vertebral separation) and spinal column partition (at the valleys of the intensity profile along the spinal column). Out of the 71 tested CT scans, only two cases had a missed partition.

The last approach described in this section consists of a spine detection followed by a classification and was done by Duong et al.[13] who try to classify spinal deformities using fuzzy clustering. They therefore create a 3D spine reconstruction in the first step using an automatic algorithm as described by Delorme et al.[10]. Next, six points per vertebra are manually extracted. As the authors state, this step can be automated by elastically registering a reference model of a vertebral body onto each of the vertebrae to be identified. After this step, the so defined features are compressed using a wavelet decomposition resulting in only 20 wavelet coefficients to describe the shape of the spine curve. These remaining features are used for a fuzzy clustering of which the output can be used for a classification. Duong et al. tested the proposed algorithm on patients with diagnosed AIS between 10 and 18 years and a measured Cobb angle greater than 40° which resulted in a maximal Rand index of 1.00 for five different classes from the fuzzy clustering, correctly separating all 409 tested samples.

## 2.2 SCOLIOSIS CLASSIFICATION

Considering the spine, one special disease is the abnormal deformation in the coronal plane which is known as (idiopathic) scoliosis. This can lead to pain in the back, increase the risk of developing shortness of breath and in many cases lowers the self-image[2]. For these reasons, it is necessary to be able to measure the severity of scoliosis and to classify these patients correctly.

The first method proposed in this section is by Thong et al.[47] and tries to classify spines of scoliosis patients. They therefore acquire biplanar X-ray scans which are used to reconstruct the spine by a semi-supervised algorithm that extracts 17 landmarks resulting in a

Figure 2.2: Fuzzy Hough Cobb Measurement **a** ROI selection **b** denoised ROIs **c** edge detection **d** edge detection without denoising

total of 867 features. These are fed into an encoder-decoder (autoencoder) neural network which leads to a reduction of features to a number of only 50 per spine. These low-dimensional codes are used for a K-Means++ clustering with five cluster centers to minimize the validity ratio and obtaining a statistical significance with $\alpha = 0.05$.

To automatically determine the Cobb angle from radiographic images, the method of Sardjono et al.[42] can be used. They use a deformable model, i.e. a modified CPM, to estimate the outline of the spine. The analogy to this model is an electric field with moving particles inside. The initial particle positions can be set manually or automatically and then move towards the left and right edges of the vertebrae by simulating the field. The Cobb angles can then easily be determined by finding the most extreme angles between the particles. Tests show that smoothing the edges defined by the particles using piece-wise linear line segments, splines or polynomials and a bias correction improve the performance and result in lower error and less time consumption (1 min to 2 min vs. 10 min to 15 min) than manually determining the Cobb angle.

Another method for determining the Cobb angle on radiographs was proposed by Zhang et al.[53]. As the first step, the image was preprocessed yielding cropped and rescaled images with the same height depicting the spinal area from C7 to sacrum. Subsequently, ROIs were manually extracted around the most tilted end-vertebrae. These patches were then denoised by anisotropic diffusion followed by an application of the Canny operator to detect the edges (see Fig. 2.2). Using this edge map, the fuzzy Hough transform was applied to extract the main direction by finding and filtering the peaks in the transformed patch. From these main directions the Cobb angle can easily be determined. To summarize this method, the mean absolute error of the results and the ground truths created by experts differed by at most 5°.

In another article, Zhang et al.[54] estimate the Cobb angle in X-ray records using a deep neural network. Like in the method proposed beforehand, the first step is a cropping and rescaling of the image but

now depicting the area from T1 to L5. 100 patches were extracted and downsampled for each shown vertebra. Those patches are the input for a DNN consisting of three hidden layers. A non-linear activation function, i.e. the hyperbolic tangent function, was only applied to the second and third hidden layer. This network was then trained to learn the slope of the vertebrae. To finally calculate the Cobb angle, the vertebrae with the largest predicted slope values were determined and their absolute values were summed up. All in all, this method produces results that differ from the true values, determined by experts, by only 5° in their mean absolute difference (like in the method before). Nevertheless, the user needs to specify the patches around the vertebrae that can then be fed into the DNN for this method as well.

Similar to Zhang et al.[53], Anitha and Prabhu[1] also proposed a method to quantify the spinal curvature using the Hough transform. At first, the radiograph is preprocessed by applying anisotropic filtering to highlight the edges. After that, the image is processed by a gradient vector flow snake (active contour) algorithm. To reduce human error, the initialization for this algorithm is determined automatically. To only retain the important edges in this segmentation, a morphological operator is applied. Finally, the Hough transform is used to determine the objective measurements. Unfortunately, Anitha and Prabhu did not provide the deviations of their method from a ground truth since their goal was to reduce the inter-observer variance.

A method for automatically measuring the axial vertebral rotation was proposed by Forsberg et al.[16]. They propose a four step algorithm that works on CT images. These steps are the extraction of the spinal canal centerline (using a circular Hough transform to find a fitting circle for the spinal canal axial-slice-wise), disc detection (by examining the intensity profile along the extracted centerline shifted towards the anterior), vertebra centerpoint estimation (by exploiting the symmetry of vertebral bodies) and vertebra rotation estimation (using the tangent of the curve running through the centerpoints). With a complete processing time of around 12 s per CT record, this method can achieve an accuracy on the same level as the inter-observer variance.

Finally, Langensiepen et al.[23] published a systematic review about methods for determining the Cobb angle automatically. After electronic searches, they found 2915 articles containing keywords like 'Cobb angle', 'computer-assisted' or 'X ray'. These were filtered drastically, yielding only 11 remaining articles in the end of the selection process. Categories were found for each of these articles to be able to classify the different methods. As a result, the following classes were identified: 'Digital-Computer assisted' (using 2, 4 or 6 landmarks or a horizontal image), 'Automatic' (using an active shape model model or the fuzzy Hough transform) and 'Smartphone' (using

Figure 2.3: Rasterstereography **a** back surface (top) asymmetry function (bottom) **b** symmetry line in lateral view (left) mean (H) and Gaussian (K) curvature (right)

the built-in accelerometer to determine the angle). As Langensiepen et al. found out, none of the methods in the articles that were classified were tested on reliability and agreement and were not compared to a reference test.

## 2.3 LORDOSIS AND KYPHOSIS

Not only is the lateral deviation of the spine unhealthy but also an abnormally excessive kyphotic or lordotic curvature of the spine. Although these types of abnormal curvature can lead to pain and other unpleasant effects as well, only few articles addressing these non-scoliotic abnormal spine curvatures have been proposed to date.

One of these methods was proposed by Reyes et al.[36] and is based on aligned RGB data and depth maps. Using a Microsoft Kinect, an image and a depth map of the patient's back are taken as the first step. These are preprocessed by applying noise removal and surface reconstruction on the depth map. Afterwards, the user can choose between a static posture analysis (SPA), a spine curvature analysis (SCA), a range of movement analysis (RMA) and a gesture recognition. The SPA computes angles and distances of 3D keypoints corresponding to dermal markers on the skin of the patient that were placed by a therapist. More interestingly, the SCA is capable of obtaining anthropometric kyphosis and lordosis. Therefore, virtual markers can be placed along the spine that are then used to generate a 3D curve using linear interpolation. Using this curve, the values of the kyphosis or lordosis curvature, respectively, can be approximated. Finally, the RMA and gesture recognition are aimed to assist in diagnoses and physical rehabilitation treatments.

A method that is also based on 3D points was proposed by Drerup and Hierholzer[12] and is able to automatically measure anatomical

landmarks on the back surface. At first, a model of the back of the patient is created by rasterstereography. After that, the symmetry line, which generalizes the medial sagittal profile to divide the back surface into a left and right part of minimal asymmetry, is extracted. This line can serve to create a reference system of the back and inversely objectively define a body-fixed coordinate system. Based on the symmetry line, many landmarks can easily be determined. Since all landmarks on one side of the symmetry line have an equivalent point on the other side, a profile of the asymmetry and a map of Gaussian curvature can be determined. This makes it easy for the user to find out the locations of deformations, e.g. caused by spine malformations. In tests, among others on patients diagnosed with lordosis and kyphosis (see Fig. 2.3 b), Drerup and Hierholzer found out that this method achieves a high accuracy and enables further measurements like the height variations of the intervertebral discs.

For treatment purposes, it is necessary to compare pre- and postoperative records of the patients for spinal malformation correction. This is why Newton et al.[32] proposed a method on biplanar radiographs to find a common reference system that they call the 'three-dimensional sagittal plane' which is able to subtract out the lateral curvature and thus keeping only the curve in the sagittal plane. This is necessary because the correction of a scoliosis can distort the radiographs in a way that the kyphosis or lordosis angles do not match any more. Newton et al. also found out that the preoperative loss of kyphosis in AIS is often underestimated because AIS inherently produces changes in the axial plane rotation.

Parvaresh et al.[33] try to predict the 3D thoracic kyphosis using 2D radiographs. They use the method of Newton et al. to compute their ground truth. To create a prediction formula for the estimated 3D T5-T12 kyphosis, biplanar radiographs as well as additional information like curve direction, apex level, Nash-Moe Grade and more are used. This formula is the result of stepwise multivariate regression of all observed variables and achieves a mean absolute error of approximately $5° \pm 4°$. During the creation of the formula, Parvaresh et al. also computed the Pearson correlation of all observed variables with the ground truth. Interestingly, only the 2D T5-T12 kyphosis was positively correlated. Almost all other variables were negatively correlated and only some were not correlated at all. Anyway, this formula was created for patients with AIS and right major thoracic curves which means it cannot be used for all patients.

# 3

# METHODS

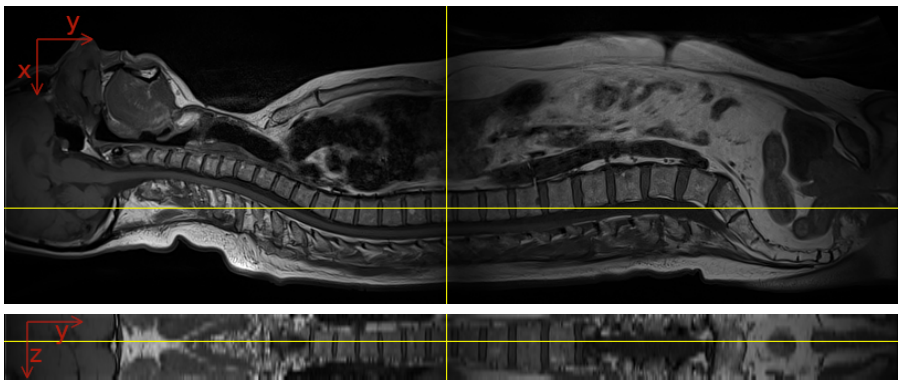


Figure 3.1: Sagittal (xy plane) slice and coronal (zy plane) slice of a record from the datasets.

In this chapter, the main methods for achieving the goal of this thesis are described.

The most important step is to extract the spine. For this purpose, many algorithms have been proposed (see Chapter 2) but all of them were either too complex (using many steps resulting in a long runtime per record) or needed further (pre-)processing (like finding regions of interest before or extracting positions from segmentations after running the algorithm). Even if many of these methods would probably have resulted in very accurate estimations of the spine, there are much simpler and faster ways to get estimations that are almost as accurate as the results of the previously mentioned algorithms.

The main objective is to obtain a comparable outcome of the method depicting the spines in a common reference system, called 'normalized spines' or 'norm spines', where MR records are the input of the algorithm. For this purpose, three datasets are available (see Section 4.1). One of these also contains the center points of the vertebral bodies from C1 to S1. Since these are given, a supervised learning method seems to be suited for the problem. One of these, and probably the most promising method as it was often applied to similar (medical) image processing problems already, is training an artificial neural network or, in this case, CNNs[27].

There are many possibilities to retrieve the spine from an MR record using CNNs. In this chapter, four of these are proposed that seem to be appropriate for the problem. Later, they are evaluated in Chapter 4 and the best performing one will be used for the statistical tests in Chapter 5.

## 3.1 STRATEGY 1

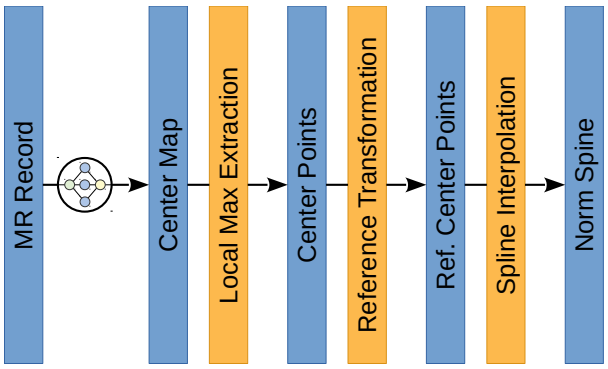


Figure 3.2: Outline of Strategy 1. Blue: outputs after the different steps. Orange: substeps. Circular icon: trainable (convolutional) neural network.

For this strategy, the first step is training a neural network that directly extracts the center points of the vertebral bodies. This seems to be straightforward yet there are some problems:

1. It is necessary for the neural network to always have the same number of output nodes. Unfortunately, the human spine does not always consist of the same number of vertebral bodies from C1 to S1 (generally 24 but in some cases more or less).

2. One could solve this by not only predicting the positions but also a confidence that shows how certain it is for some vertebra to exist. One could choose a high number of centers to be predicted (preferably as high as the maximal number of vertebrae that can ever be found) and let the network decide how many vertebrae it thinks there are in the input record. Anyway, this would presumably not be very robust since it adds one more degree of freedom per vertebra prediction and the order of the vertebrae in the predicted output can be variable.

3. (Fully) convolutional neural networks are known for not sufficiently perform regression tasks like this one, even if fully connected layers are put behind the convolutional layers[14], but rather to do classifications (or segmentations).

For these reasons, the regression problem is mapped to a classification/segmentation problem in the following way: The output of the network will not be the coordinates of the center points of the vertebrae but a matrix with the same size as the input with ones at the positions of a vertebra center and zeros elsewhere. Since the vertebra centers are distributed very sparsely compared to the rest of the elements of the matrix, it would be almost impossible for a neural

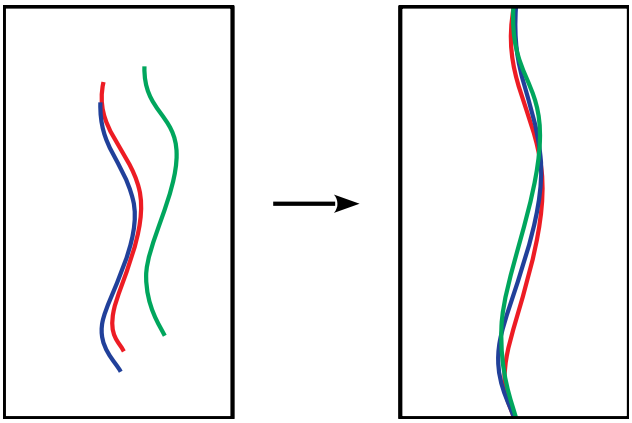

Figure 3.3: Exemplary transformation of three spines (left) into the reference system (right). Notice that all norm spines have the same start and end point in the reference system.

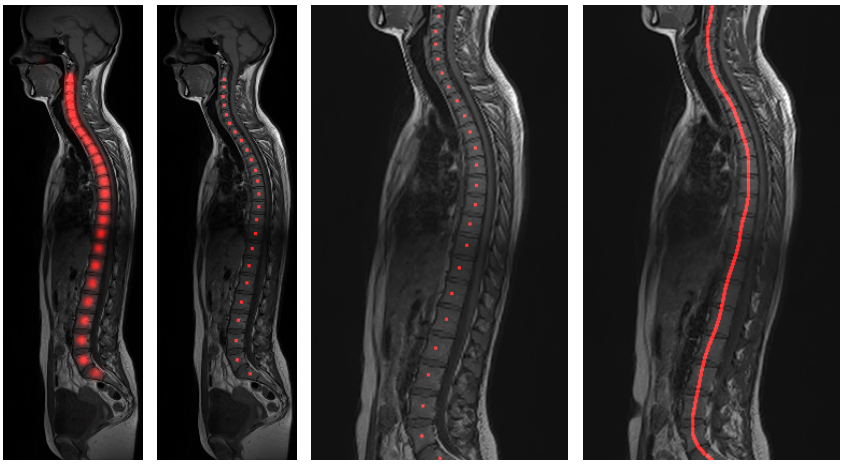

Figure 3.4: Z-Projections of midsagittal slice of intermediate outputs of Strategy 1. F.l.t.r.: center map, center points, reference center points, norm spine

network to learn these points. That is why a Gaussian filter with a certain kernel size is applied to the so-created matrix. This causes values around the center points to get a value unequal to zero as well, which can be seen as a kind of probability density function. Nevertheless, the smoothing by applying the Gaussian kernel causes the values at the center points to not being one any more. Since the task is still a kind of segmentation, these values should be one anyway, which is why the matrix is normalized by dividing by its maximum, such that the minimal value stays at zero but the maximal value is mapped to one and all values in between are scaled properly.

After training the network to create the described outputs, which will be called 'center maps' from now, it is still necessary to extract the coordinates of the center points. Assuming a perfect center map, i.e. a generated map for training the network, it should be easy to find the searched points by using a simple maximum finder. Therefore, a

sliding window approach can be used. The window size should be small enough to contain only one maximum, i.e. vertebra center, but should not only contain one voxel. A voxel can then be classified as vertebra center if its intensity is higher than every other intensity in the window. Unfortunately, the predictions of the network will not be perfect which will cause other artificial maxima, especially where intensities are close to zero, to be found by this method. To work around this problem, a threshold can be introduced that controls the minimal intensity for a voxel to be classified as a real maximum.

Assuming the found maxima correspond to the true centers of the vertebral bodies, the next step is now to map the points into a common reference system (see Figure 3.3). This can be done by applying a similarity transform (being able to rotate, scale and translate the input without reflections) to each of the center points which transforms them such that the topmost point $p$ will be mapped to the reference point $r_t$ and the lowest point $q$ will be mapped to the reference point $r_l$, respectively (for a detailed description of how to compose this transformation, see Appendix A.1).

To map all centers into the reference system, the so-found transformation must be applied to all centers.

The output should be the curve of the spine and not only the centers of the vertebrae which is why some interpolation is needed for the values in between the center points. In this case, a simple cubic spline interpolation is used to achieve this[44].

## 3.2 STRATEGY 2

The first step of this method is to extract the centerline of the spine directly from the MR record by using a CNN. The ground truths for training this network can be created by using the given centers of the vertebral bodies. Therefore, a cubic spline is fit through the centers. Again, as mentioned in Section 3.1, a CNN is more efficient in learning a segmentation than a regression. For this reason, the cubic spline is embedded inside a matrix of the size of the input Magnetic Resonance Image (MRI), setting all values along the centerline, i.e. the spline from the uppermost vertebra center to lowest center, to one and everything else to zero. Yet again, a Gaussian kernel is used to smoothen the values of the matrix, since the values are still very sparsely distributed. This smoothed centerline will be called 'center ridge' or 'ridge map'.

Parallel to this CNN extracting the center ridge, another CNN is trained to learn the center of the uppermost and the lowest vertebra. The ground truths for this CNN can be created by using the centers of the vertebrae as well. In this case, the output of the CNN is going to have two channels, one for the upper vertebra center and another one for the lower one. To create these channels, the list of all centers

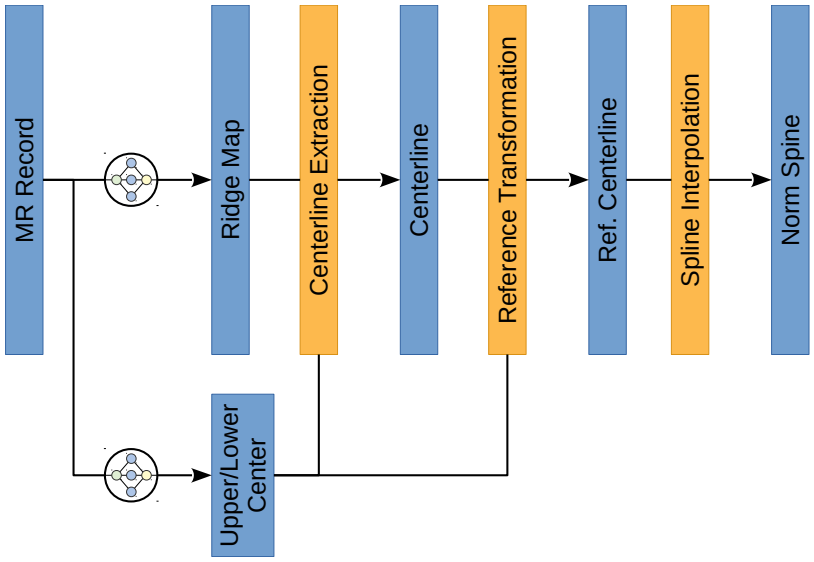


Figure 3.5: Outline of Strategy 2. Blue: outputs after the different steps. Orange: substeps. Circular icon: trainable (convolutional) neural network.

of the MR record is sorted by the y-value, i.e. the height. For the first channel, the first center of the list is used and the last center of the list is used for the second channel, respectively. This assumes that there are no two centers in the list having the same y-value which should be given by the nature of the spine. Now, for each channel, a zero-matrix with the size of the input MRI is created, only being one at the position of the respective center. To adopt the idea of center point extraction in Section 3.1, a Gaussian kernel is applied to each channel to help learning the segmentation. It would also be possible to combine both centers in one channel, as it was done with all centers in the previous strategy. However, it would be necessary to search for local maxima which can lead to wrong points (as pointed out in the previous strategy). The benefit of using two separate channels for each of the centers is that only one global maximum needs to be found per channel, which is much less complex than a local search and yet should lead to precise coordinates. Furthermore, the uppermost and the lowest vertebrae are used as reference points because they have a characteristic appearance that makes it easier and robust to learn by a CNN.

The next step is to extract the centerline. Therefore, the coordinates of the maximum of the x-z-plane (i.e. an axial slice) at every height of the center ridge is searched. Since the coordinates of the uppermost and the lowest vertebra center are known from the second CNN, it is only necessary to extract the maxima between these two points. All other points above and below the centerline will be discarded, anyway. It can be possible for the extracted centerline to not be very smooth since the maxima at every height were extracted

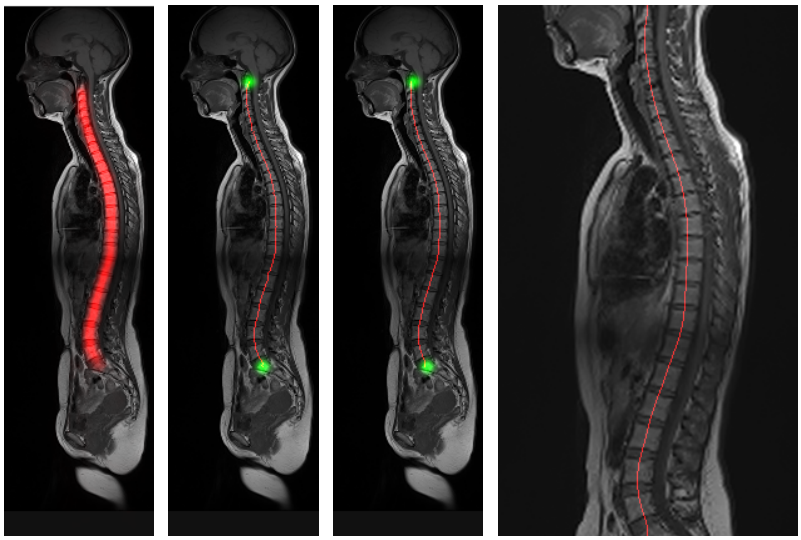

Figure 3.6: Z-Projections on midsagittal slice of intermediate outputs of Strategy 2. F.l.t.r.: ridge map, centerline (red) + upper/lower center(green), smoothed centerline + upper/lower center, norm spine

independently from each other. For this reason, it seems suitable to apply a Gaussian filter to the extracted points. This one should have a small kernel because otherwise the curvature of the spine could be distorted.

Of course, it would be possible to use more robust algorithms for extracting the centerline from the center ridge like applying a shortest path algorithm with an appropriate cost function. This would probably be more robust than finding the maximum at every height level because it considers previous and following height levels, but as can be seen later in Chapter 4, this simple method is sufficiently robust and very fast.

At this point, the coordinates of the extracted spine are still in the original system. This is why they need to be projected into the reference system which can be done by applying the similarity transformation from Section 3.1 to all extracted points where the parameters for calculating the transformation can be taken from the output of the second CNN, i.e. the coordinates of the uppermost and the lowest vertebra center.

The last step is to interpolate missing values of the spine in the reference system which can occur due to the similarity transformation of the points from the original system. Therefore and once again, a cubic spline is fit through the current points of the spine and at every height, an interpolated value can now be obtained.

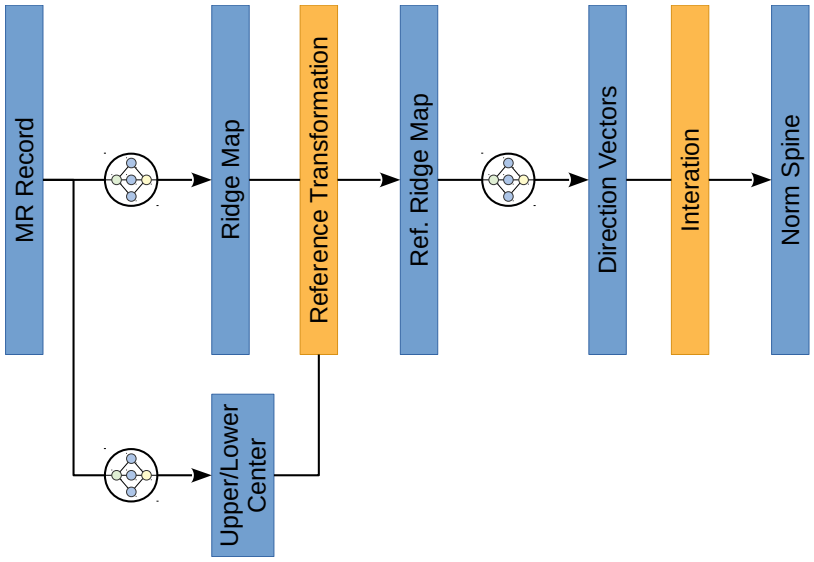


Figure 3.7: Outline of Strategy 3. Blue: outputs after the different steps. Orange: substeps. Circular icon: trainable (convolutional) neural network.

## 3.3 STRATEGY 3

This strategy is very similar to the preceding one in the first steps but takes another approach of extracting the spine in the reference system.

Such like in Strategy 2, the input MRI is first fed into a CNN which outputs the center ridge. A second CNN which extracts the center of the uppermost and the lowest vertebral body is used as well.

This is the point where the strategies start to differ. Instead of extracting the centerline from the center ridge, a similarity transformation is applied to the center ridge that transfers the ridge into the reference system. The parameters of this transformation can be obtained in the same way as it was presented in Section 3.1 but it is necessary to invert the so-created matrix since in this case, the coordinate system is transformed instead of single points[1]. A third CNN is trained with these normalized center ridges as input and will output a three-dimensional (normalized) direction vector at every height. Each of these direction vectors points down along the normalized spine. Even if this is a regression task, one can assume that this can work well because the direction vectors are similar to gradients, i.e. vectors of partial derivatives, and it is easy for CNNs to train a filter that can compute gradients based on applying a convolution with an appropriate kernel (the simplest kernels being the Prewitt or Sobel operator[34, 46]).

1 Instead of calculating the inverse matrix, it is more efficient to swap $p$ with $r_t$ and $q$ with $r_l$ and use this matrix directly.

The final step is now to integrate the direction vectors to retrieve the true positions. This consists of the following substeps:

1. Forward-Integration: Let $d_i$ be the normalized direction vector at height $i$. $d_{i,y}$ denotes the y-component of $d_i$. The true position $t_i$ at height $i$ can then be computed using the recursive formula

$$t_{i+1} = t_i + \frac{1}{d_{i,y}} \cdot d_i \,, \qquad 1 \leq i < k$$
$$t_1 = r_t.$$

2. Backward-Integration: The following recursive formula results in the same points $t_i$ for a perfect estimation of normalized direction vectors $d_i$:

$$t_i = t_{i+1} - \frac{1}{d_{i,y}} \cdot d_i \,, \qquad 1 \leq i < k$$
$$t_k = r_l.$$

3. Let $t_i^f$ denote the points from the forward- and $t_i^b$ the points from the backward-integration. Using a perfect estimation of normalized direction vectors, it holds true that $t_1^f = t_1^b = r_t$ and $t_k^f = t_k^b = r_l$ for both the forward- and backward-integration. Since there will be small errors on the estimations due to the prediction of the CNN, it is likely to happen that $t_k^f \neq r_l$ and $t_1^b \neq r_t$. Anyway, it still holds that $t_1^f = r_t$ and $t_k^b = r_l$. To combine this and to assure both $t_1 = r_t$ and $t_k = r_l$ for both interpolations, it is possible to interpolate the values linearly. Then the final estimation of the points will be:

$$t_i = \frac{k-i}{k-1} \cdot t_i^f + \frac{i-1}{k-1} \cdot t_i^b \,, \quad 1 \leq i \leq k.$$

## 3.4 STRATEGY 4

The last strategy that is going to be tested is a rather direct method of extracting the normalized spine in the reference system.

Once again, a CNN is trained with the MR records as input. Unlike the previous strategies, this CNN will not output an intermediate result like the center ridge but at every height the coordinates of the centerline, i.e. a regression problem. Therefore, it is certainly important to create the centerline, first. This can be done like in the previous strategies by fitting a cubic spline to the given center points of the vertebral bodies. Since it should be easier to train the network such that it outputs values around zero (no linear bias and no additional scaling need to be learned), the coordinates of the points on the centerline are normalized to be in the interval $[-1, 1]$. This needs

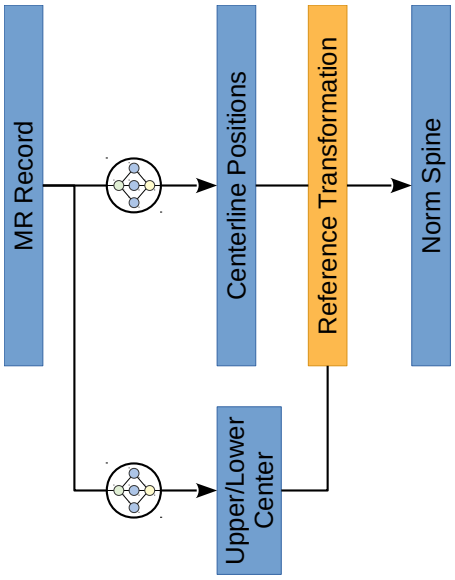


Figure 3.8: Outline of Strategy 4. Blue: outputs after the different steps. Orange: substeps. Circular icon: trainable (convolutional) neural network.

to be done for both the width and the depth coordinate which results in the mappings $[1, width] \mapsto [-1, 1]$ and $[1, height] \mapsto [-1, 1]$. Of course, it is necessary to set some artificial values at height levels where there is no centerline. Some strategies to set these points are:

1. Set the coordinates to zero, i.e. the center of the axial slices.
2. Set the coordinates above the centerline to the coordinates of the uppermost point of the centerline and the coordinates below it to the coordinates of the lowest point, respectively.
3. Set the coordinates to values outside the defined interval of $[-1, 1]$.

At first glance, it seems unnecessary to specify these strategies because values above and below the centerline will not be needed for the normalized spine, anyway. Nevertheless, setting these points to certain coordinates can drastically impair the training process and thus deteriorate the predicted coordinates of the actual centerline.

After the coordinates of the points on the centerline have been extracted using the CNN, it is necessary to reverse the scaling/translation of the width and depth dimensions such that the values of the coordinates will be in $[1, width]$ and $[1, depth]$ again. It would be possible to not reverse this transformation but then the output of this strategy would be different from the outputs of the other three strategies.

The last step is to transform the extracted centerline into the reference system by applying the similarity transform from Section 3.1 to

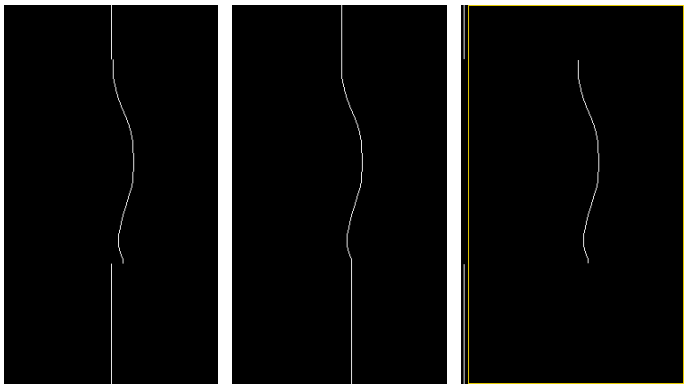

Figure 3.9: Visualization of strategies to set points outside the spine for Strategy 4. Left: zero. Center: continue upper/lower points. Right: outside $[-1, 1]$ (yellow border delimits the image area; notice the white lines left of the area)

each of the points on the centerline. Therefore, it is necessary to determine the uppermost and the lowest point of the centerline which can be done by predicting the center points of the uppermost and lowest vertebra using the CNN proposed in Section 3.2. Like in the previous strategies, it is important to interpolate the missing values after the transformation by using a cubic spline.

# 4 EXPERIMENTS AND EVALUATION

The methods proposed in Chapter 3 need to be tested and evaluated to verify their quality. In this chapter, several tests on the different steps of the method are carried out and discussed in terms of the applicability to the problem.

## 4.1 DATASETS

### 4.1.1 *SHIP Study*

The data that the proposed method is tested on has been created during the Study of Health in Pomerania (SHIP)[49] which is a population-based epidemiological study carried out in the north-eastern part of Germany (see Figure 4.1) on several thousand subjects since 1997. Its goal is to estimate the prevalence of different diseases and to find connections between simultaneous occurrences of multiple diseases. The different datasets built up in the study are:

- SHIP-0: data from patients recorded from 1997 to 2001
- SHIP-1: 5-year follow-up examination of patients from SHIP-0 between 2002 and 2006
- SHIP-2: 11-year follow-up examination of patients from SHIP-0 between 2008 and 2012
- SHIP-Trend-0: new cohort recorded from 2008 to 2012
- SHIP-3: 17-year follow-up examination of patients from SHIP-0 between 2014 and 2016

Figure 4.1: Region of SHIP Study in Germany

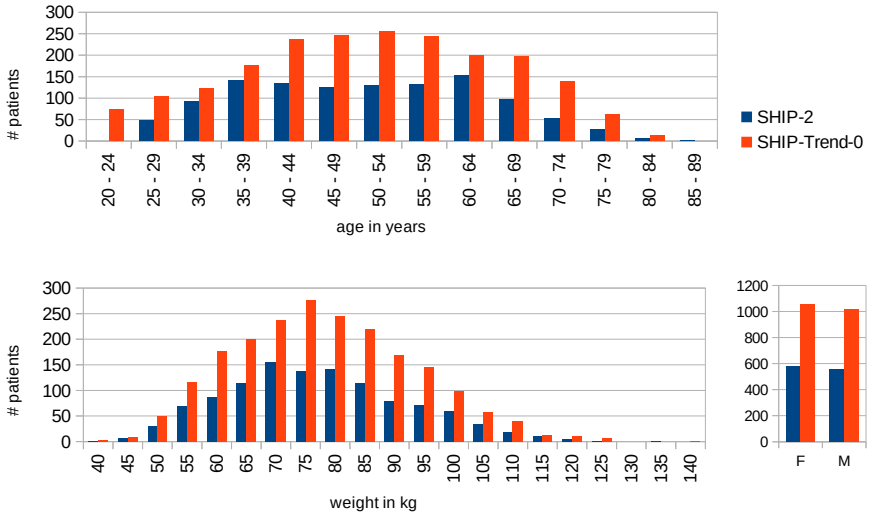


Figure 4.2: General statistics of SHIP-Study: distribution of age (top), weight (bottom left) and sex (bottom right)

- SHIP-Trend-1: follow-up examination of patients from SHIP-Trend-0 since 2016

The data gathered in each of the datasets consists not only of interviews about the health of the patients (containing questions about many different diseases like diabetes or asthma) and laboratory data (information about DNA, blood cell count, calcium, magnesium and many more) but also medical examinations reaching from blood pressure measurements, ultrasound records of the liver and dental examinations to dermatological examinations, sleep monitoring and whole-body MRI.

To preferably get a uniform distribution of the population, a stratification based on region, age and sex has been performed for the choice of subjects participating in the study. From Figure 4.2, the stratification can be verified for sex with a ratio of 1639 female to 1571 male subjects. A stratification of age can also be seen for SHIP-2 from 35 years to 64 years (binned in ranges of five years) with each approximately 125 subjects. There do exist people with ages outside this range in the SHIP-2 study but they are not stratified. In contrast to that, the age of the subjects in SHIP-Trend-0 seems to be a Gaussian distribution with the mean value at the age range from 50 years to 54 years. As can be seen in the bottom left graph of Figure 4.2, the stratification was not applied to weight which again results in a rather Gaussian distributed random variable with its maximum at 75 kg to 79 kg (the labels in the graph show the lower value of the weight range binned in 5 kg steps).

There is one more dataset called 'SHIP-Pretest' which was conducted on a small number of volunteers that are not contained in any of the other datasets. This served to detect problems in the pro-

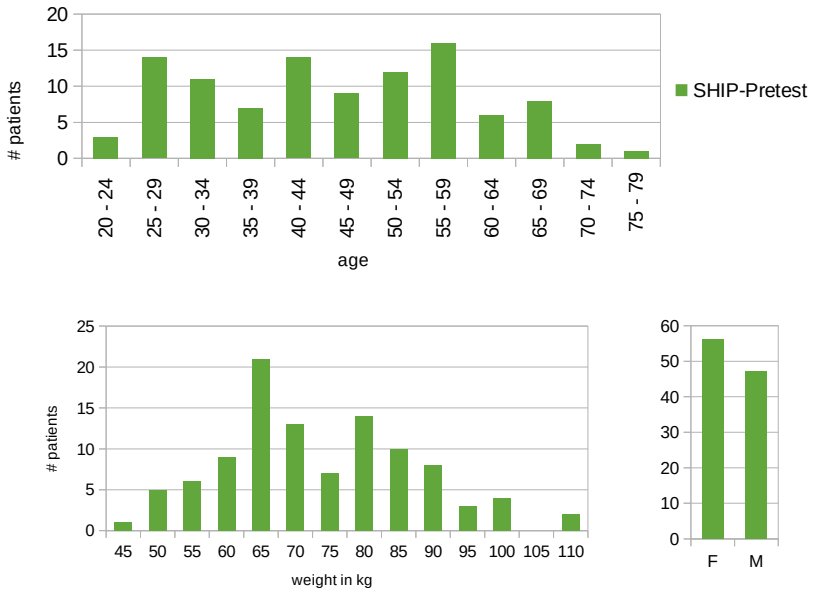


Figure 4.3: General statistics of SHIP-Pretest: distribution of age (top), weight (bottom left) and sex (bottom right)

cess of gathering the data before the actual study. As can be seen in Figure 4.3, there has not been applied stratification to any of the features age, weight or sex. Nevertheless, the sex seems to be balanced with a ratio of 56 female to 47 male subjects and even the age distribution seems to be rather uniform between the ages 25 and 59 (again binned into intervals of 5 years) with approximately 10 subjects each. For the weights, there does not seem to be a simple distribution but it can be assumed to be a Gaussian-distributed random variable similar to the weight distribution in the two big datasets before. Even if none of the three features seems to be sampled from a uniform distribution, they show a great variance of values which will later be beneficial for training the neural networks.

In this thesis, only the MR records of the sets 'SHIP-Pretest' (containing 103 patients), 'SHIP-Trend-0' (containing 2072 patients) and 'SHIP-2' (containing 1138 patients) will be used where 'SHIP-Pretest' will be used for training the networks and the other two datasets will be used for computing the statistics. It was not possible to use scans from 'SHIP-0' or 'SHIP-1' because the imaging procedure first started in 'SHIP-2'.

Each subject was scanned using a 1.5 T MR Imager (Magnetom Avanto; Siemens Medical Systems, Erlangen, Germany) in the supine position where phased-array surface coils were situated on the head, neck, abdomen, pelvis and the lower extremities, and the spinal coil was placed inside the patient table. These scans resulted in both T1-weighted and T2-weighted records per patient. Each sagittal slice has a field of view of approximately 50 cm $\times$ 50 cm with a pixel spacing of

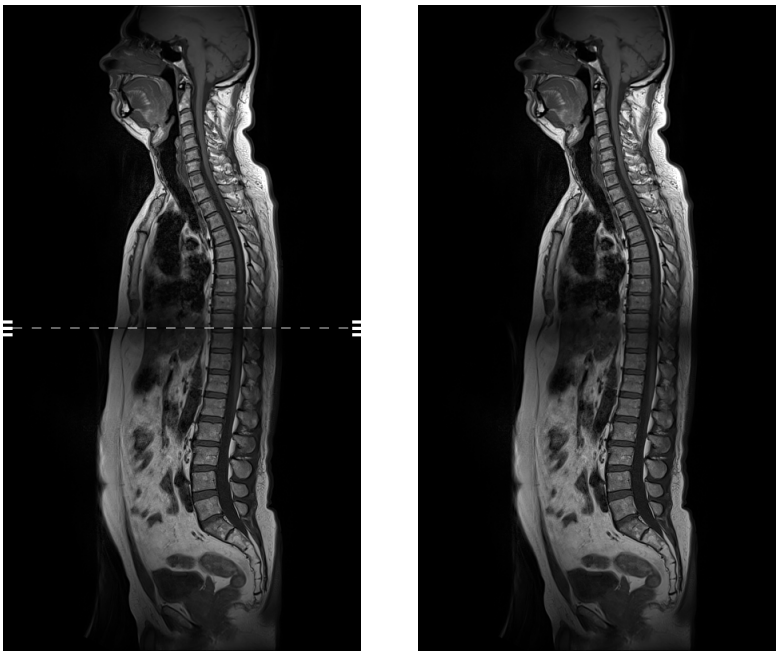

Figure 4.4: Midsagittal T1-weighted MR slice unprocessed (left) and processed (right)

1.116 mm × 1.116 mm. The resolution along the sagittal axis is lower with a spacing between the slices of 4.4 mm resulting in a thickness of around 7 cm to 8 cm for the whole record. Since the field of view of the MRI was quadratic, it was not possible to record the whole body in one shot. Rather, the area used in this thesis, containing the spine, was visible in two consecutive shots which had to be combined (see Figure 4.4). In these combined records, markers for alignment were contained as well. For the reconstruction of the whole body, these markers had to be removed at first and were instead filled up with interpolations of the missing values. This had to be done for both the T1-weighted and the T2-weighted records.

### 4.1.2 *Preprocessing and Data Augmentation*

As Ioffe and Szegedy[19] found out, it is useful, i.e. enables a faster and more robust learning, for training neural networks to normalize the input data, i.e. subtracting its mean and dividing by its standard deviation. This process can be done inside the network on mini-batches of inputs on the activations of arbitrary (preferably all) layers, which allows to remove the (trainable) bias vectors of the convolutional or dense layers, or also (as it was done in this case) on each input MR record independently.

Still, the dimensions of the so-created records could differ by some voxels which would make the use in neural networks impossible without further processing. For this reason, the maximal values for each

dimension of the dataset were found and each normalized record was embedded in a matrix of the found maximal dimensions. The rest of the matrix was filled up with zeros (i.e. the mean value). Since the resolution of the provided MR records was much higher in the sagittal plane than along the sagittal axis and to accelerate training, the dimensions of the sagittal plane were halved resulting in a final shape of $240 \times 425 \times 17$ of the network input matrix.

It is also helpful to create additional artificial data from the existing data by applying transformations like random flippings, rotations, scalings and translations as proposed by Ciresan et al.[8] since the number of possible input records from the training set of 103 patients is very small. In this case and only during training, the matrices could be flipped along the sagittal axis (left-right flip) or along the transversal axis (up-down flip) and they could be translated along the coronal axis (front-back translation) or the transversal axis (up-down translation). Each of these four transformations could take place with a probability of 50 %. The values for the translations were found by the following procedure: Since it should be helpful for training to have a wide variety of samples, the translations should be maximal in extreme cases, i.e. it is only necessary to keep the centers of the vertebral bodies inside the boundaries of the matrix. To achieve this, the minimal and maximal coordinates of all spine centers of all training subjects were found and random values for the front-back/up-down translation were computed for every augmented sample using these minimal and maximal coordinates such that the vertebrae centers would never leave the matrix.

### 4.1.3 *Ground Truth Creation*

Since neural networks are trained iteratively by backpropagation and thus are one of the supervised machine learning algorithms, it is necessary to create a set of outputs for the given inputs for training. It is also necessary to have data representing the desired outputs of the method to be able to validate the accuracy and correctness. This data is called *ground truth*. For the purpose of this thesis, it is necessary to know the centers of the vertebrae to be able to estimate the spine curvature. The SHIP-Study does not include any information about the spine or the vertebrae which is why the centers of the vertebral bodies had to be extracted manually. In this case, only the vertebrae from C1 to S1 have been extracted since these represent the biggest part of the spine, are well-visible in the records and look very similar to each other (opposing to e.g. S2-S5). Because the resolution of the original MR images along the sagittal axis is approximately four times lower than in-plane, it would often have been the case that the positioning of the center points along the sagittal axis would have been too coarse (even visually). For this reason, the resolution was

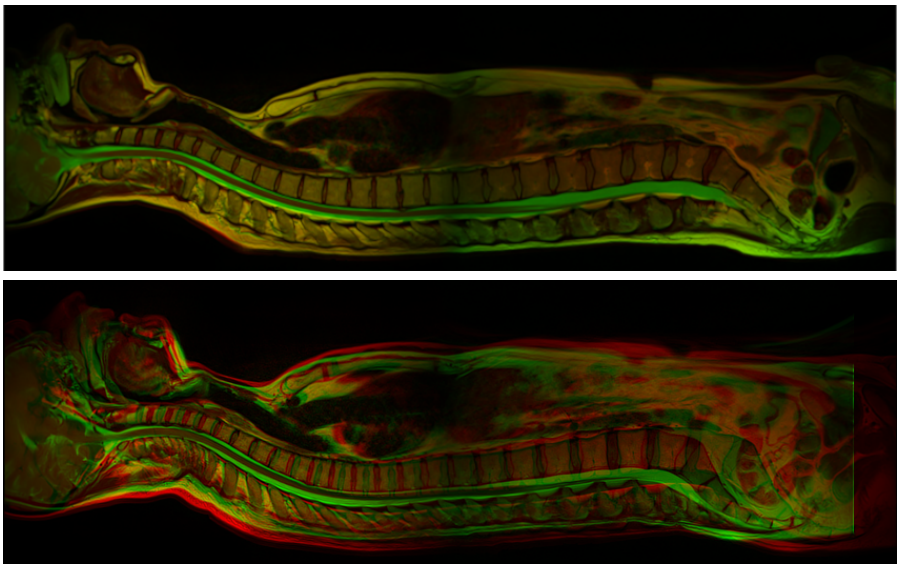

Figure 4.5: Aligned (top) and misaligned (bottom) T1- and T2-weighted records. Red channel: T1, green channel: T2

increased artificially during ground truth creation by inserting three slices between two consecutive slices in the record resulting in a new sagittal resolution of 65 instead of 17 where

$$R_i^H = R_j^L, \quad j = \lfloor i/4 \rfloor, \quad i \in [0,64], \quad j \in [0,16]$$

for the high resolution record $R^H$ and the low resolution record $R^L$. Using this magnification, it is not only possible to achieve sub-voxel precision but also obtaining a record whose in-plane sagittal pixel spacing is almost equal to the spacing between the slices. It is important to mention that the centers of the vertebrae have been created using only the T1-weighted records. This is because in some cases, the T1-weighted MR record and the T2-weighted record were not aligned properly (see Figure 4.5) or had slightly different resolutions which would have distorted the coordinates of the extracted centers. Using the so-obtained center points of the vertebrae, it is possible to create the desired outputs of the networks to carry out the training process.

## 4.2 EVALUATION OF THE STRATEGIES

In this section, the implementation of the different strategies proposed in Chapter 3 is described. After that, the strategies are evaluated and their applicability to other datasets (in this case 'SHIP-2' and 'SHIP-Trend-0') is discussed.

### 4.2.1 *Strategy 1*

Briefly explained, this strategy takes an MR record as input and predicts the center map from one CNN and the uppermost and lowest vertebra center from a second CNN. The center points are extracted from the center map by a local maxima search and they are transformed into the reference system by using the output of the second CNN. Finally, a cubic spline is fit through the transformed centers.

The only free parameters of this strategy are the two CNNs and the choice of the reference system. The system is, like for all strategies, set such that the center of the uppermost vertebra is at $(120, 0, 32)$ and the lowest one at $(120, 424, 32)$. These values could be set arbitrarily but like this, the extracted spines will only be transformed slightly and a discretization of the interpolated values will still be very accurate. As the transformation changes the spacings of the voxels, the units of the reference system will be called *rpx* (for 'reference pixel').

As for the networks, they need to fulfill a (fuzzy) segmentation task. In recent years, one architecture or adaptations of this architecture are the state of the art in medical image analysis and outperform all previous methods[11, 25, 26]. This architecture is called *U-Net* and was proposed by Ronneberger et al.[40] in 2015. It is a kind of auto encoder network, i.e. consists of an encoding path, which consists of alternating convolutional- and pooling layers, and a decoding path, that consists of alternating upsampling- and convolutional layers. What makes the U-Net so powerful are connections between the encoding and decoding path at every subsampled layer which means that small details from early layers can be used in later layers. For these reasons, the networks used for this strategy are adaptations of the U-Net as well. Their structure can be taken from Figure 4.6.

As can be seen, the architectures of the two CNNs are very similar and only differ by the parameters of individual layers. The encoding path comprises three-dimensional convolutions with a kernel size (KS) of $3 \times 3 \times 3$ and a number of filters of 16, 32, 64 and 128 for the first CNN and 8, 8, 16 and 16 for the second one. Each convolutional layer is followed by a batch normalization layer, a nonlinear relu activation and a max pooling layer on certain axes (except for the last convolutional layer of the encoding path).

The decoding path consists of alternating upsampling layers on certain axes followed by a concatenation with the results of the encoding path at the same sampling level which are then followed by convolutional layers, again with a KS of $3 \times 3 \times 3$, with 64, 32 and 16 filters for the first and 16, 8 and 8 filters for the second network (again with batch normalization layers and relu activations). The last layer carries out a convolution with a KS of $1 \times 1 \times 1$ and one filter

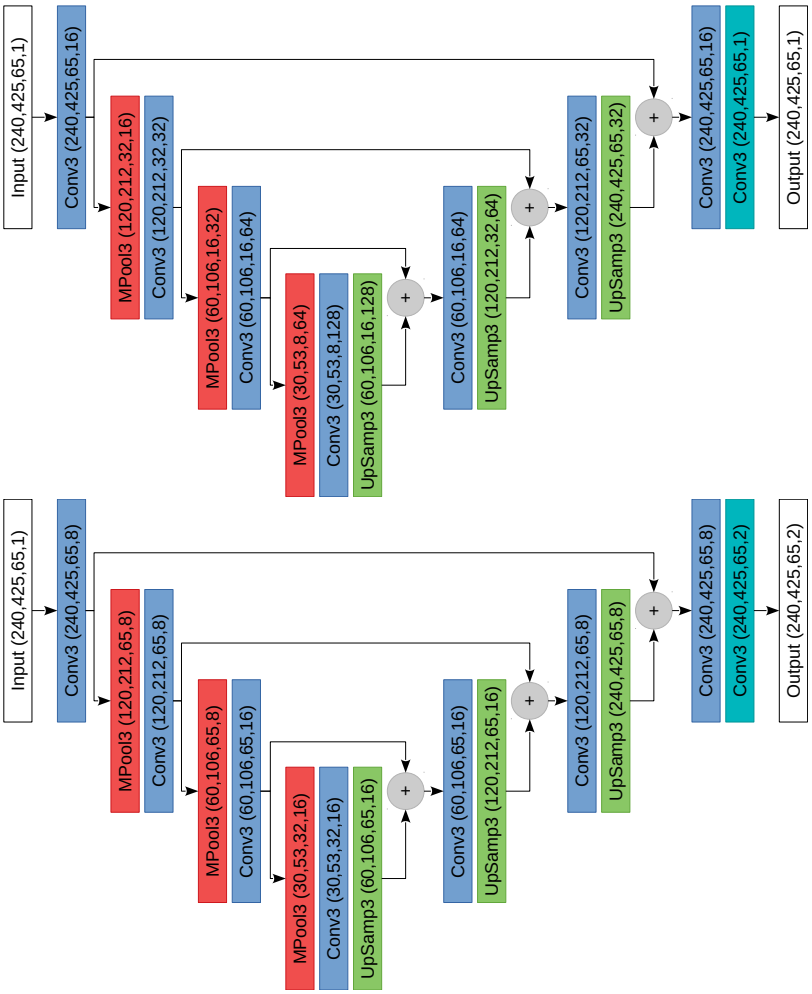


Figure 4.6: CNNs of Strategy 1. Top: U-Net extracting center map. Bottom: U-Net extracting upper and lower vertebra center. Blue: convolution+relu, red: max pooling, green: upsampling, gray circle: concatenation+zero padding, turquoise: convolution+sigmoid. Values in brackets: output shape of the respective layer

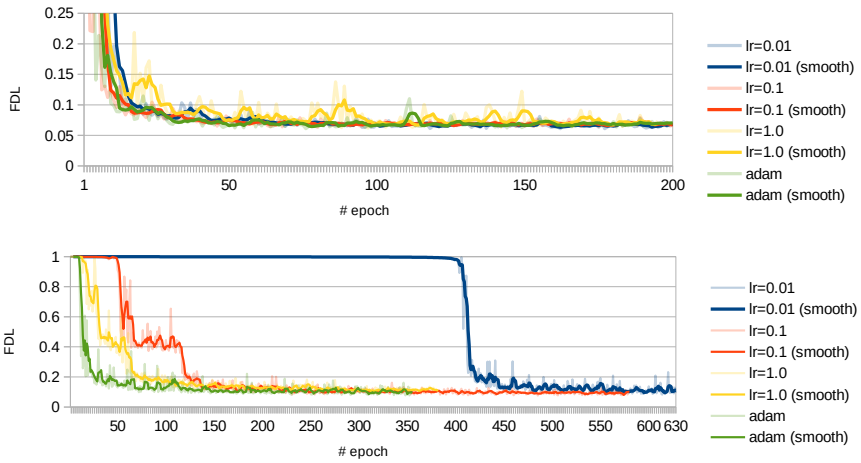


Figure 4.7: Losses of Strategy 1. Top: Center map extraction net. Bottom: Uppermost/lowest vertebra extraction net. Losses using SGD (learning rates: 0.01, 0.1, 1.0) and Adam

for the first and two filters for the second CNN which output the final predictions.

For training the two networks, the Fuzzy Dice Loss (FDL) function was used as objective function to be minimized. It is defined as

$$FDL = 1 - \frac{k + \sum_i |g_i \cdot p_i|}{k + \sum_i g_i^2 + \sum_i p_i^2},$$

where $g_i$ and $p_i$ are the intensities of the ground truth and the prediction, respectively, at position $i$, and $k$ is a smoothing constant[30].

In all cases, a batch size of 2 was used for training and 1 for validation. The networks were trained as many epochs as were necessary for them to converge. In this case, a network converged when it trained for at least 200 epochs and reached an FDL below 0.2.

The optimizer, that is used to train the networks, is SGD[38] with a learning rate of 0.01, using Nesterov's Accelerated Gradient[31] and a momentum term[35] of 0.7. These values were found by a coarse grid search of the parameters to result in smooth learning curves and to simultaneously achieve a small value of the loss function.

Figure 4.7 shows the graphs of the validation loss of several parameter configurations. The first graph shows the FDL of the network that learns the center map. One can see that the loss decreases very fast in the first ten epochs to approximately 10 %. In the next 150 epochs, it falls much slower to a minimal value of 6 %. As one can see, increasing the learning rate does not significantly change the minimum of the loss curves but makes them rather uneven. Even changing the optimizer from SGD to Adam (using its default parameters) does not have any impact on the losses, at least for this network.

The second graph shows the FDL of the network that learns the uppermost and lowest vertebra center with different parameter con-

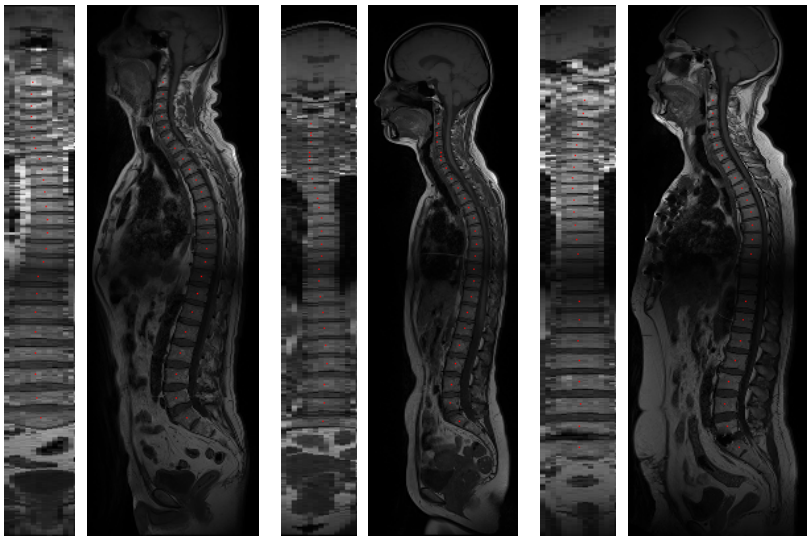

Figure 4.8: Predicted centers of Strategy 1. Left: Correct prediction. Center: Too many predictions (see cervical area). Right: Too few predictions (see dark thoracic area)

figurations. As can be seen, the configuration with the default parameters described in the previous paragraph needs 400 epochs before it falls from 98 % to 20 % in approx. 15 epochs. If the learning rate is increased to 0.1 or even 1, the loss decreases much faster, i.e. reaches 20 % after 120 or 70 epochs, respectively. On the other hand, the minimal values of the loss get larger the more the learning rate is increased: the minima of the learning rates 0.01, 0.1 and 1 are 8 %, 8 % and 9 %, respectively. This means that an optimal learning strategy must be a tradeoff between speed and accuracy, at least for this network. This tradeoff does not only depend on the learning rate of SGD but on the optimizer, too. Comparing all graphs of SGD with the one that used Adam, one can see that SGD is slower but can achieve slightly better losses than Adam which is why SGD was used for further tests.

In Figure 4.8, three examples of the extracted centers from the predicted center maps can be seen, embedded in a sagittal CPR of the input MRI. The left one shows an accurate prediction whereas the central one shows an erroneous one with 26 detected centers and the right one shows an erroneous one with 22 detected centers. As it turns out, if 26 centers are detected, then the two additional centers are either situated in the cervical area where the vertebrae look similar to intervertebral disks due to the low resolution, or in the thoracic region where the original upper and lower record are connected and the illumination is bad, or inside S1 whose size is bigger compared to the other vertebrae and thus the window of the local minima search is probably too small. However, this only occurs in 2.5 % of all cases on the datasets 'SHIP-2' and 'SHIP-Trend-0'. In case of only 22 detected centers, the problem always lies in the bad illumi-

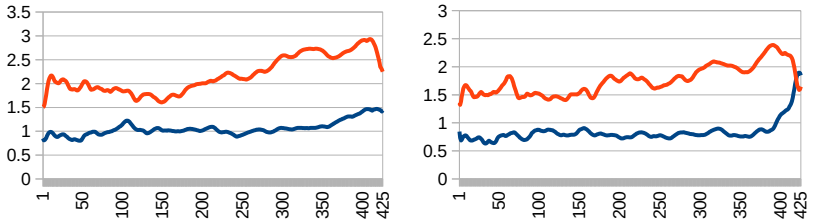


Figure 4.9: Absolute error of Strategy 1. Left: Mean absolute error per height level for x (blue) and z coordinate (red). Right: Standard deviation of errors per height level for x (blue) and z coordinate (red).

nation between the upper and lower record and occurs in even less cases, i.e. 0.4 %, on the two big datasets. If the method outputs 24 centers, then they are correct in all cases. Regarding the other two cases, i.e. 23 and 25 predicted centers, 20 % and 60 %, respectively, miss at least one center because of illumination issues or contain at least one center too much, which makes 1.2 % and 10.6 % of the whole dataset. All in all, the probability that the method does not output the correct number of vertebra centers is 14.7 %.

Nevertheless, it is not too important to gather the correct number of centers but rather a cubic spline from these centers whose error to the cubic spline of the real centers is small. This was tested and is depicted in Figure 4.9. What can be seen is the mean absolute error (MAE) and the standard deviation (STD) at every height level between the spline interpolated with the ground truth centers and the centers predicted by this strategy for anteroposterior (x) and lateral (z) coordinates. To make things shorter, these values come from the reference system, i.e. the final output of this strategy. One can see that the MAE of the x values is never greater than 1.5 rpx (approx. 3.3 mm) whereas the error of z values is at least 1.5 rpx (approx. 1.65 mm) and at most 3 rpx (approx. 3.3 mm). This can probably be explained by the (originally) lower resolution along the sagittal axis. The results are similar for the STD yet often up to 0.5 rpx less than the corresponding MAE for both x and z values. This shows that the error is relatively constant throughout the datasets.

The second network, i.e. for predicting the center of the uppermost and the lowest vertebra, was also tested and the results are shown in Figure 4.10. The boxplot shows that it seems to be easier for the network to predict the position of the uppermost vertebra center opposing to the lowest one: The median for the upper center is at 1.22 rpx compared to 2.1 rpx for the lower one. Even if the first and the third quartile of the two centers are rather close to each other, the maximal error of the lower center is 5.24 rpx compared to only 4.06 rpx for the upper one. This means that the distance between the calculated and the true center coordinate is between 1 rpx and 2 rpx to 2.5 rpx in most cases but occasionally, the calculation of the lower center can differ from the true one quite a bit. In the binned distribu-

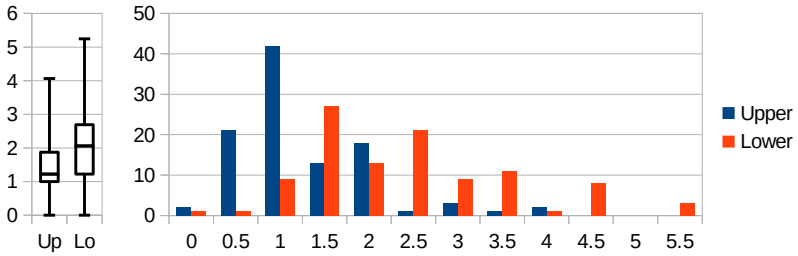


Figure 4.10: Accuracy of Network 2 in Strategy 1. Left: Boxplot of MAE of uppermost and lowest vertebra center. Right: Binned distribution of MAE of the respective centers.

tion, one can see an interesting deviation from an assumed normal distribution at bin 2 for the lower center which contains less values than the previous and the next bin. One explanation for this could be the choice of bin sizes or the fact that only discrete positions can be output.

### 4.2.2 *Strategy 2*

This method is similar to Strategy 1 but the first network does not predict a center map but a ridge map of which the ridge is extracted, normalized and smoothed.

Since the prediction of the ridge map is a (fuzzy) segmentation task as well, a U-Net-like architecture is chosen such like in Strategy 1 (see Figure 4.6). In fact, the exact same architecture, i.e. with the same layers and the same filter properties, is used but trained such that it outputs the ridge maps from the MRI. The loss curves are very similar to those of the network in the previous method, which is why they are not shown here.

The second network, i.e. for extracting the uppermost and the lowest center of all vertebrae, is the same as in the previous method as well and was not trained again, so even the same weights have been used for testing this strategy. This also means that the same properties of this network still hold.

After the maxima of the ridge map were extracted at every height level and the so-found ridge positions have been transformed to the reference system, it is useful to smooth the points because of the reasons named in Section 3.2. This adds one more free parameter to this strategy, i.e. the size of the Gaussian filter kernel. In Figure 4.11, the absolute errors between the calculated and the true normalized spine can be seen for three different smoothing KSs. The left graph shows the error of anteroposterior (x) values whereas the right one shows the error of lateral (z) values. It can be seen that the error is less than 2 rpx at almost all height levels for the x coordinate. Only at the last 14 levels (in the sacral region), the error exceeds the error

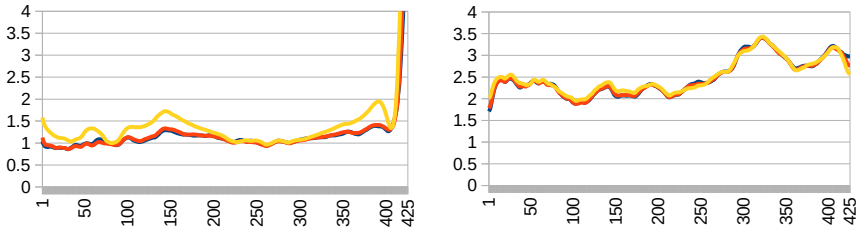


Figure 4.11: Accuracy of Strategy 2. Left: Absolute error of x values for different kernel sizes (KS). Right: Absolute error of z values for different KSs. Blue: KS=3. Red: KS=5. Yellow: KS=10.

of 2 rpx and ends at 8 rpx, 9 rpx and 12 rpx for KSs of 3 rpx, 5 rpx and 10 rpx, respectively (which cannot be seen on the graph). For the z values, the errors are mostly between 2 rpx and 3 rpx which can be explained by the low resolution along the sagittal axis and the artificial magnification described in Section 4.1.3. Nevertheless, even if the sagittal axis was enlarged by a factor of 4, the error is only two to three times as large as along the coronal axis. It is also interesting to see that the error does not get much larger at the last height levels, which means that it it easier to predict the lateral curvature (possibly because it is rather straight, normally). The values of the standard deviation are not shown here but are very similar to those of the MAE (see Appendix A.1). Concerning the different KSs, it can be stated that the error gets larger the larger the kernel gets, which proves that the assumption made in Section 3.2 holds true. Regarding the MAE, its value for the x errors is minimal for a KS of 3 (i.e. 1.20 rpx compared to 1.22 rpx and 1.45 rpx for KS = 5 and KS = 10, respectively) whereas its value for the z errors is minimal for a KS of 5 (i.e. 2.48 rpx compared to 2.49 rpx and 2.52 rpx for KS = 3 and KS = 5, respectively). Since KS = 3 and KS = 5 produce similar errors but KS = 5 is visually more pleasant and more natural, all further tests will use KS = 5.

### 4.2.3 *Strategy 3*

In this strategy, the ridge map extracted by the same network as in the previous strategy is transformed into the reference system using the second network from the previous strategy and a third CNN estimates direction vectors along the ridge (i.e. the spine) at every height level which are finally integrated to get the estimated spine positions.

The network for estimating the direction vectors is the only one, described in this thesis, that does not have an architecture similar to a U-Net but rather a straightforward CNN using only alternating convolutional layers and poolings (see Figure 4.12). After the activation of every convolutional layer (except for the last one), a batch normalization is carried out. As soon as a dimension reaches 1 after a

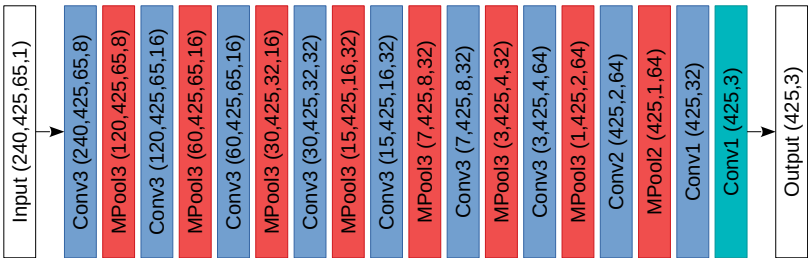


Figure 4.12: CNN for direction vector extraction. Blue: convolution+relu, red: max pooling, turquoise: convolution+linear activation. Values in brackets: output shape of the respective layer

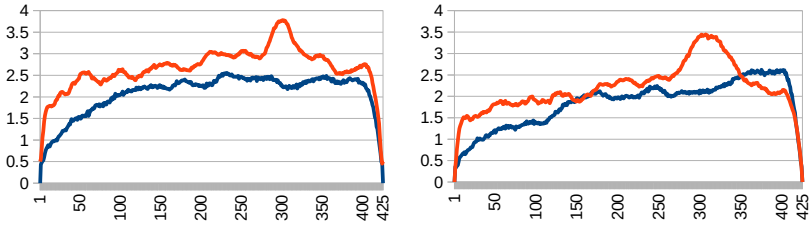


Figure 4.13: Absolute error of Strategy 3. Left: Mean absolute error per height level for x (blue) and z coordinate (red). Right: Standard deviation of errors per height level for x (blue) and z coordinate (red).

pooling, the matrix is reshaped to have one dimension less. That way, all further convolutional layers can be lower-dimensional and thus less parameters need to be learned, which enables a faster training process.

In Figure 4.13, the absolute errors of the calculated output from the true norm spine can be seen as well as their standard deviation for x and z values along the y axis. Different from all other proposed methods, the absolute error and the standard deviation reach zero at certain height levels, namely at the start and at the end. This is obvious in this case because of the way this strategy calculates the norm spine: Since the coordinates are interpolated integrals from direction vectors, the start and end must always be correct. This is why it is more important to look at the values in between. One can see that both x and z errors increase to 2.5 rpx. The z error increases even more to 3 rpx until it reaches a short peak of 3.8 rpx at height level 302 (i.e. in the lower thoracic region). This peak could be explained by the strong curvature of the spine in this region and the relatively wide shape of the vertebral body at that position. During the last 20 height levels (i.e. the sacral region), the error decreases very fast until it reaches zero.

The standard deviation looks very similar. It also starts and ends in zero and has a peak in the z values in the transition zone between the thoracic and lumbar spine (at height levels around 300). Nevertheless, all values are approximately 0.5 rpx less than their respective absolute errors at the same height level.

### 4.2.4 *Strategy 4*

The last strategy directly extracts the coordinates of the points along the spine using a CNN. Like in most of the previous methods, a second CNN is run in parallel which detects the uppermost and the lowest vertebral body. These are then used to transform the extracted coordinates of the first CNN into the reference system.

As pointed out in Section 3.4, there are several different ways to handle points above and below the actual spine. Tests showed that using the method that assigns 0 to the coordinates of these points works best. After all, the results are still not satisfactory enough to be used for an estimation of the spine curve as can be seen in Figure 4.14 (left): The positions are very uneven, even if the CNN allows for smoothing them. Furthermore, the values at the bounds of the actual spine are estimated very badly and especially uneven. The reason for this could be that the values after the bounds jump to 0 and an exact border cannot be determined by the network.

A slight modification of this method has been tested as well: Instead of directly extracting the positions using the CNN, the ridge map is extracted first and mapped to the reference system. The advantage is that choosing a method for setting values outside the spine can be omitted since these do not exist any more. Using calculated ground truths of the normalized ridge as input for the position detection CNN still results in rather uneven estimations as in Figure 4.14 (center). Since the small kernel that is used in the CNN layers could be too small to be able to smooth the positions, another version with kernel sizes of 25 has been tested as well, even if it is not usable well in a real application due to its slow training. A result can be seen in Figure 4.14 (right), which shows a rather smooth spine but still has some irregularities.

Since the results are not even close to the ground truth norm spines visually, a calculation of errors and their standard deviations was waved.

## 4.3 CHOICE OF THE STRATEGY

Comparing all four strategies, it can be concluded that Strategy 3 and Strategy 4 produce the largest errors or are not even usable in a real applications. For these reasons, the focus is on the first two strategies.

Comparing the MAEs of them, it can be stated that Strategy 1 produces errors around 1.0 rpx for x and 0.8 rpx for z whereas Strategy 2 produces errors around 1.0 rpx and 2.5 rpx, respectively. In contrast to that, concerning the standard deviation, Strategy 1 has values around 2.5 rpx for x and 2.0 rpx for z whereas for Strategy 2, the values are around 0.8 rpx and 2.0 rpx, respectively. Since Strategy 1 creates a

Figure 4.14: Results of Strategy 4. Prediction using 0 outside the spine (left). Prediction using norm ridge as input (center). Prediction using large kernels (right). For visualization, the z coordinates of the positions have been omitted and the remaining points have been drawn onto a zero matrix, after which the matrix was smoothed (which results in a ridge map).

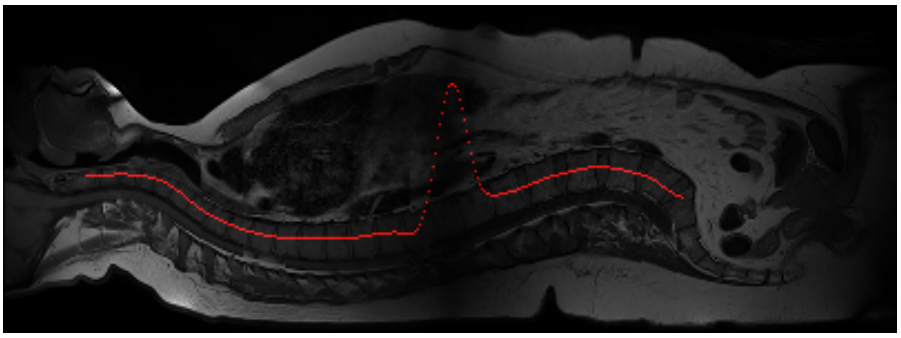

Figure 4.15: Erroneous result of Strategy 2

wrong number of vertebra centers in 14.7 % of all cases and the standard deviation of errors of Strategy 2 is less, the choice fell on Strategy 2.

It is important to note that there can be cases where the ridge is not predicted completely correctly by the network in this strategy, although this only happened in ten cases that all look like the one in Figure 4.15. This can probably be explained by the bad illumination in the center of the records.

# 5 STATISTICAL EVALUATION

In Chapter 3, some strategies for extracting the spine from an MR image and mapping them into a common reference system were presented and the best one has been chosen in Chapter 4. In this chapter, several statistical tests on the so-extracted spines are carried out.

## 5.1 VISUALIZATION

For a better and more local interpretation of the results in the normalized coordinate system, the spine is visually divided into three parts depicting the cervical, thoracic and lumbar part of the normalized spine. Therefore, the two bounds, one between the cervical and thoracic part and one between the thoracic and the lumbar part, have to be found. Since the centers of the vertebral bodies exist for the dataset 'SHIP-Pretest', it is used for calculating these two bounds. For the upper one, the following steps are carried out:

1. For each record, the centers corresponding to C7 (the lowest cervical vertebral body) and Th1 (the uppermost thoracic vertebral body) are found by sorting all centers by their y-coordinate, i.e. their position along the transversal axis.

2. For each record $r$, the transformation for the normalization of the spine is calculated as an affine matrix $A_r$ like in Section 3.1 using the coordinates of the centers of C1 and S1, respectively.

3. For each record $r$, the centers of C7 and Th1 found by Step 1 are transformed into the normalized coordinate system using matrix $A_r$ from Step 2.

Figure 5.1: Visualization of all norm spines. Top: sagittal projection. Bottom: coronal projection. White: centerline. Light gray: 1st to 3rd quartile. Dark gray: 5th to 95th percentile. Green: cervical area. Violet: thoracic area. Yellow: lumbar area.

| Vertebra | Mean | Std | 5th Perc | 95th Perc | Deviation |
|---|---|---|---|---|---|
| C7 | 78.4 | 3.1 | 74.0 | 84.0 | $[-4.4, 5.6]$ |
| Th1 | 92.7 | 3.2 | 88.5 | 99.0 | $[-4.2, 6.3]$ |
| Th12 | 295.0 | 5.2 | 290.1 | 301.3 | $[-4.9, 6.3]$ |
| L1 | 319.9 | 5.4 | 314.2 | 325.6 | $[-5.7, 5.7]$ |

Table 5.1: Simple statistical measures of certain normalized vertebra centers: mean, standard deviation, 5th percentile, 95th percentile, 90% data deviation from mean in rpx

4. The mean of all normalized centers of C7 is calculated as $\mu_{C7}$.
5. The mean of all normalized centers of Th1 is calculated as $\mu_{Th1}$.
6. Since the border between the cervical and the thoracic region should be approximately the center between C7 and Th1, the mean of $\mu_{C7}$ and $\mu_{Th1}$ is calculated and only the y-coordinate is taken.

For the lower bound, all of these steps can be carried out likewise, but C7 is replaced with Th12 and Th1 is replaced with L1.

To check that it is appropriate and meaningful to use the mean of the normalized centers in Step 4 and Step 5, i.e. to find out whether the deviation from the mean is small, one can calculate the interval that most of the values are inside of. This can be seen in Figure 5.1 where for each necessary vertebra the mean, the standard deviation, the 5th percentile, the 95th percentile and the interval of deviation from the mean is shown. The interval of deviation $[a, b]$ means that 90 % of the data is in the interval $[\mu + a, \mu + b]$ for the mean value $\mu$ where $\mu + a$ is the 5th and $\mu + b$ is the 95th percentile, respectively. As one can see, the majority of values does not deviate more than 6.3 rpx from the mean value. It can also be derived that the deviation gets (a little) larger the lower the position of the vertebral body is. It is also interesting to see that the mean is not centrally located in the interval of deviation in most cases. Instead, the distance to the 5th percentile is less than to the 95th percentile. Since the deviation is at most 6.3 rpx and the 95th percentile of a vertebra is not greater than the 5th percentile of the next lower vertebra, it can be concluded that the mean is a good estimation for the center positions.

Finally, the coordinate of the border between the cervical and the thoracic region in the normal system is at $0.5 \cdot (78.4 + 92.7) = 85.6$ and the one between the thoracic and lumbar region at $0.5 \cdot (295.0 + 319.9) = 307.4$. The regions can be seen in Figure 5.1.

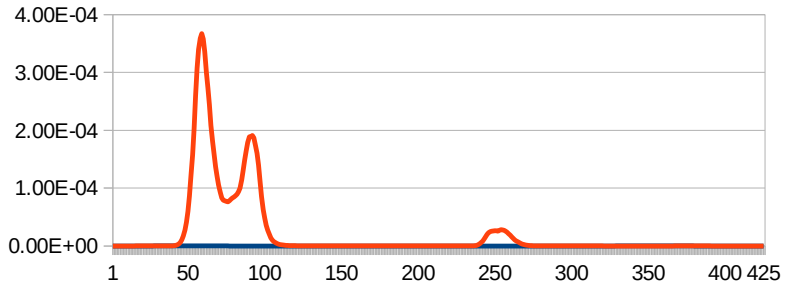


Figure 5.2: Shapiro-Wilk test: p-values at every height level for x (blue) and z values (red)

## 5.2 TEST FOR NORMALITY AND MODALITY

The first test that is carried out on extracted norm spines is a test on normality, i.e. whether a random sample is drawn from normal distribution. The intention of this test is to find out whether further tests can be used that assume a normal distribution (like Student's t-test). Else, different methods must be used that do not assume normality.

The method used for checking the normality is the Shapiro-Wilk test[45] in this case. There are other possible methods but this one is suitable since it supports sample sizes up to 5000 and the datasets to be tested contain approx. 4000 samples. The null-hypothesis of this test is that the samples were drawn from a normal distribution. Since the input of this test can only be one-dimensional data, the method was carried out for every height level of the norm spine and for anteroposterior and lateral direction (x and z values) separately.

Figure 5.2 shows the p-values for x and z coordinates at every height level which describes the probability to draw the samples like this assuming the samples were drawn from a normal distribution. Since the maximal p-value is $3.67\mathrm{e}{-4}$, it means that it is very unlikely that the samples were drawn from a normal distribution. In most cases, especially for the x values, the p-values is even smaller and does not even reach $5\mathrm{e}{-5}$. From this, it can be concluded that the norm spines are not (or very unlikely) normally distributed.

It is also interesting to know the number of modes of the norm spines. If it is greater than $1$, it could be possible that there are several clusters that describe different shapes of the spine. Since the test for normality failed, the upper limit of one mode cannot be sustained. For this reason, Hartigan's Dip Test [18] is carried out which assumes the unimodality of the tested samples as its null-hypothesis. Figure 5.3 shows the results of the Dip Test. In this case, the closer to 0 the value is, the more likely it is for the samples to be drawn from a non-unimodal distribution. Setting the common significance level $\alpha = 0.05$ would at almost all height levels not allow to reject the hypothesis of a non-unimodality. Most of the height levels have

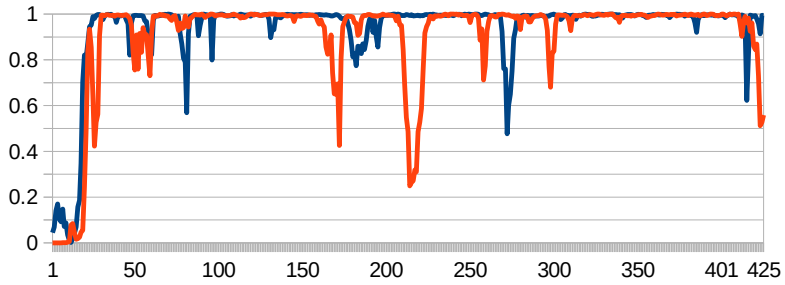


Figure 5.3: Hartigan's Dip Test: p-values at every height level for x (blue) and z values (red)

a p-value very close to 1 which indicates very likely that the samples were drawn from a unimodal distribution.

Summarizing these two tests, the norm spines are not normally distributed and they also probably do not have more than one mode on most height levels.

## 5.3 TEST ON STRAIGHT CORONAL SPINE

It is widely assumed that an ideal spine does not have a lateral deviation. Real patients, however, do not have ideal spines and thus, a small lateral deviation is rather possible. Nevertheless, the hypothesis to be checked is that the mean of all spines in the given datasets is a straight line (in a coronal view).

Therefore, the wanted test needs to compare the norm spines at every height with the required expectation value, i.e. the center in the reference system. A test that is capable of doing so is the Wilcoxon signed-rank test which checks whether the distribution of the differences is symmetric about zero. The results can be seen in Figure 5.4 (top).

One can see that in most cases, the p-value is very close to zero which means that the hypothesis that the means coincide with the straight line must be rejected. This shows that the mean coronal norm spine is not a straight line but rather slightly curved. This can also be seen visually in Figure 5.1 (bottom).

Another method to show that the mean spine is not completely straight is by calculating the length of the spines in the coronal view which should optimally be 424 rpx (corresponding to the length of a straight line in the reference system from C1 to S1). The distribution of these lengths is shown in Figure 5.4 (bottom). One can see that most lengths are between 424 rpx and 429 rpx. The mean of 426 rpx also confirms that a typical spine is longer than the optimal value what can only occur if the spine is not straight.

Finally, one can conclude that the initial hypothesis, that the mean coronal spine is straight, is not true.

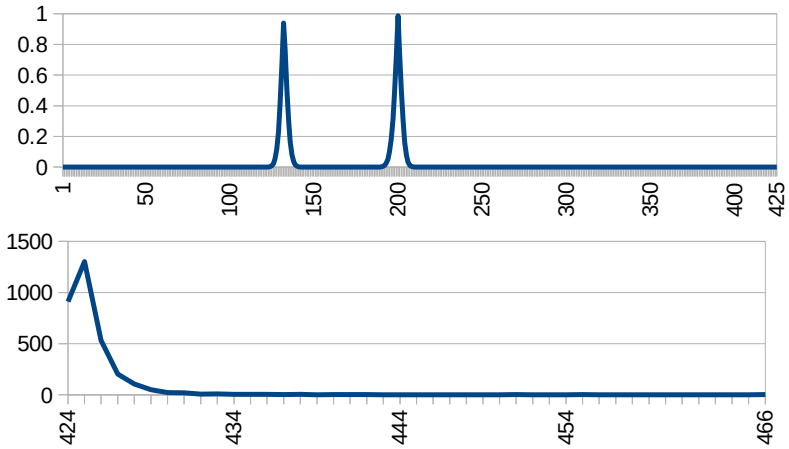


Figure 5.4: Test on Straight Coronal Spine. Top: p-values of Wilcoxon signed-rank test comparing the dataset to the expected straight line at every height level. Bottom: binned histogram of coronal spine lengths (50 bins with a size of 1 rpx each)

## 5.4 FILTERED TESTS

As previous analyses on the curvature of the spine found out[15, 17], the shape of the spine is likely to change with age, body size, weight and other features. For this reason, diagnosing diseases of a patient's spine based on the statistics of spines of the whole dataset can be misleading. E.g., comparing the spine of an old person with the mean spine of the population will probably show a strong deviation and thus lead to the diagnosis of a disease like hyperkyphosis although compared to the mean spine of people at the patient's age would only show a mild deviation. This is why the whole set of spines will be filtered for certain features in this section such that tests between different groups of people become possible.

### 5.4.1 *Test on Biological Sex*

The feature that is most constant in the life of a human being is the biological sex, i.e. male and female. Since men and women do develop differently in anatomy but still not as different as it should concern the spine, the hypothesis is that the estimated probability density functions of male and female subjects do not differ.

To validate this hypothesis, a Kolmogorov-Smirnov (KS) test is carried out between the male and female spines, which assumes all samples to be drawn from the same distribution. The results can be seen in Figure 5.5 (bottom) which shows the p-values of the test. In this case, if the p-value is high, the hypothesis cannot be rejected. One can see that the values are close to zero for the anteroposterior direction (along the x axis) in most cases except for the transition zone between the thoracic and lumbar spine (around height level 300). It

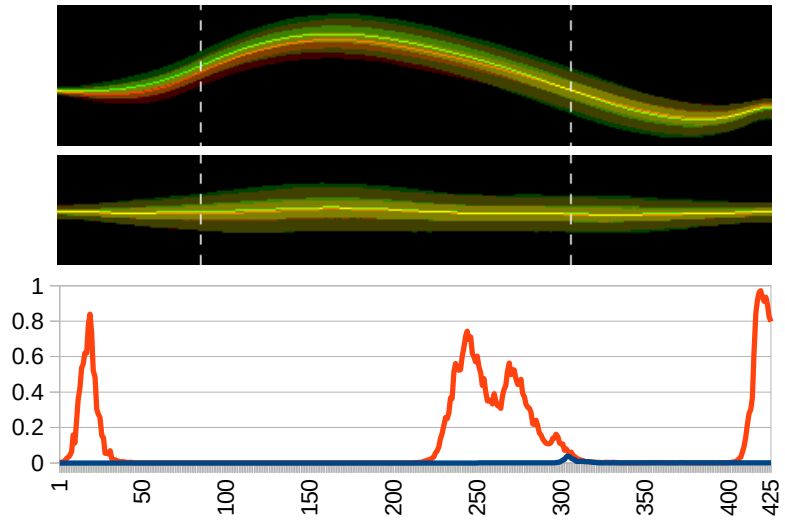


Figure 5.5: Test: Male vs. Female. Top: Sagittal projection of male (red channel) and female spines (green channel) using 5th, 25th, 75th and 95th percentile and mean. Center: Coronal projection. Bottom: p-values of KS test at every height level (blue: x coordinates, red: z coordinates)

is possible to set the common significance level at $\alpha = 0.05$. All p-values are below $\alpha$ which shows the significance of this test for the anteroposterior direction (i.e. the x values). Applying the same $\alpha$ to the z values only holds true for 71 % of the height levels. This is because the lateral curvature is not as strong as in the sagittal plane and lets the distributions at some height levels overlap. Nevertheless, even for the z values, the test is significant in most cases.

All in all, this shows that male and female spines do differ which means that the stated hypothesis is wrong. This can also be seen in Figure 5.5 (top, center) where differences can be seen especially in the cervical and thoracic area.

### 5.4.2 *Test on Body Size*

The body size is another feature that can have an impact on the curvature of the spine. To check for this influence, the patients were divided into size bins of 10 cm from 1.3 m to 2 m. The distribution can be seen in Figure 5.6 (top). One can see that it is shaped like a normal distribution with its mean between 1.6 m and 1.7 m. Comparing the bins visually (as for the bins 1.5 m and 1.8 m in Figure 5.6 (above center)) already shows a difference: The norm spines of large patients are closer to the center than those of small patients, i.e. large patients seem to have a straighter spine.

To validate this thesis, a Mann-Whitney-U (MWU) test is performed on the height bins which can check for the alternatives 'smaller', 'unequal' and 'greater'. Since picking only two bins for the validation

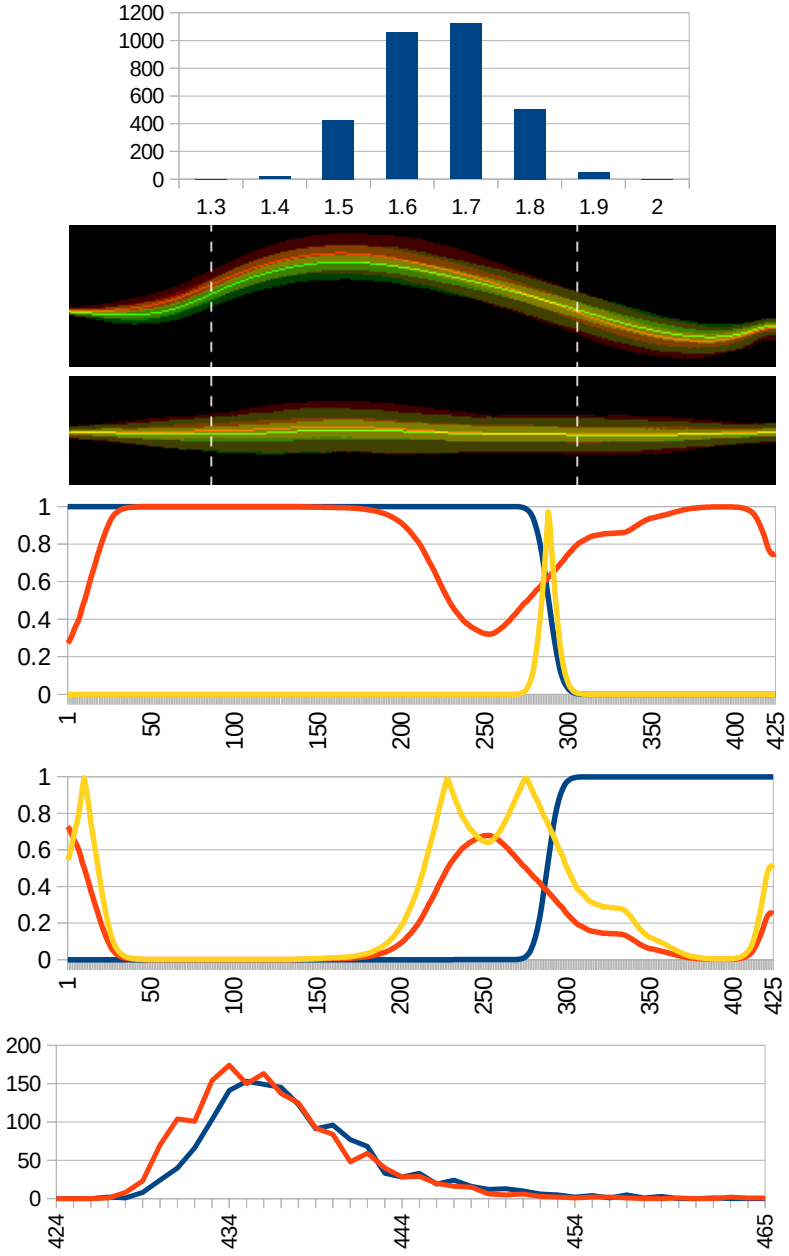


Figure 5.6: Test on Body Size. Top: distribution of sizes. Above center: sagittal and coronal projection of norm spines of bin 1.5 (red channel) and 1.8 (green channel) using 5th, 25th, 75th and 95th percentile and mean. Center: p-values of MWU test for x values per height level for the alternatives 'less' (blue), 'unequal' (yellow) and 'greater' (red). Below center: like lower center but for z values. Bottom: Binned distribution (1 rpx per bin) of sagittal lengths of norm spines for size$<$1.7 m (blue) and size$\geq$1.7 m (red).

would not be sufficient and a test on all combinations would be too costly, a critical size was set and the test will compare all norm spines of sizes below to those of at least that critical value. The mean of the distribution lies between 1.6 m and 1.7 m, so it should be sensible to use 1.7 m as the critical value (see Figure A.3 and Section A.3 for a further justification of this value).

For the anteroposterior direction (x values), one can see that the inequality can be assumed in most cases except for the lower thoracic area. This is the place where the norm spines have a zero-crossing (in the reference system) and for that reason are superimposed. For all values in the lumbar region, the alternative that the x values of the smaller bin are less than those of the larger bin can be assumed as well, which indicates that the larger spine is straighter in this region. In the cervical and thoracic region, none of the alternatives 'less' or 'greater' can be assumed although visually, the larger bin seems to be straighter. More directly, comparing the sagittal lengths of the norm spines of these two sets (see Figure 5.6 (bottom)) visually shows a distribution of larger spines being shifted towards smaller lengths. A modified version of the sign test (see Appendix A.3) on the lengths using the alternative 'greater' also confirms this with a p-value of $1.5\text{e}{-10}$.

For the z values, a significant statement can only be made in the lower lumbar region where the alternatives 'unequal' and 'greater' can be assumed. In all other regions, the test cannot determine the order relation between the two bins. In the upper thoracic region, the test even assumes all three alternatives. One can suspect that this happens because an assumption of the test could possibly be not fulfilled: MWU assumes that the distributions are the same but can be shifted. Testing the bins for equal but shifted distributions would go too far which is why it is not done here.

Finally, the hypothesis that spines of larger people are straighter can be accepted visually and (for the x values) statistically.

### 5.4.3 *Test on Age*

The age of a patient plays a prominent role for the shape of the spine as it was shown i.a. by Fon et al.[15] who showed that the mean kyphosis increases with age. So the hypothesis is that spines of younger people are straighter than those of older people. Visually (see Figure 5.7 (top)), it can be determined that the norm spines of younger patients are indeed straighter than those of older people, at least in the sagittal plane. Furthermore, one can see that the variance of younger spines is smaller that that of older spines.

An MWU test cannot be performed on this data because the variances of the distributions to be compared differ which violates the assumption of equal but shifted distributions. Instead, a KS test is

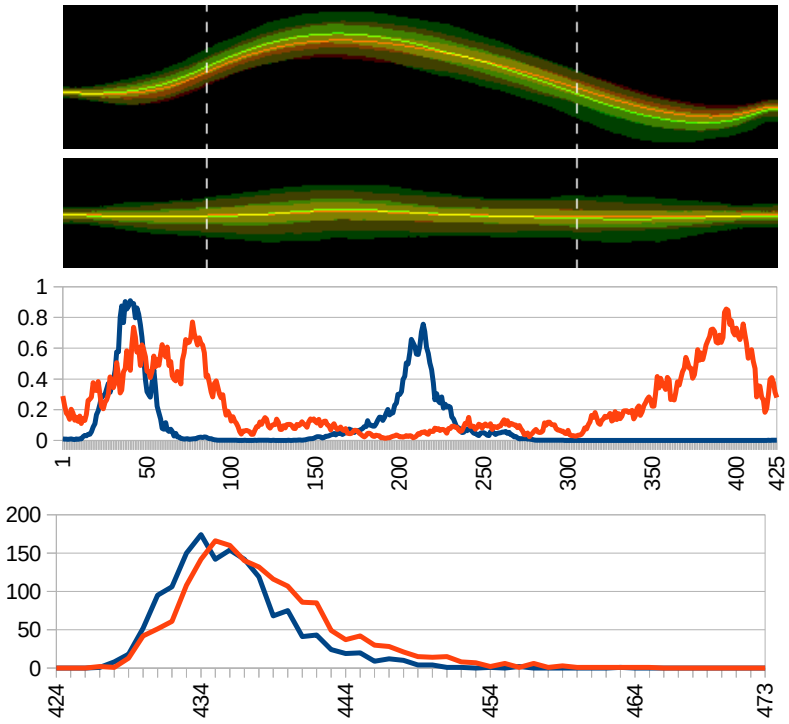


Figure 5.7: Test on Age. Top: Sagittal and coronal projection of norm spines of bin 20 (red channel) and 70 (green channel) using 5th, 25th, 75th and 95th percentile and mean. Center: p-values of KS test for x (blue) and z (red) values per height level. Bottom: Binned distribution (1 rpx per bin) of sagittal lengths of norm spines for age$<$45 years (blue) and age$\geq$45 years (red).

performed (see Figure 5.7 (bottom)). Like in the previous subsection, the test is not performed on two single age bins. Rather, all norm spines smaller and greater than a critical age are combined in two sets. This value is set to 50, since the weighted mean of the distribution lies between the bins 40 and 50. As previously assumed, the distributions of the x values differ in most areas except in the lower thorax at the inflection point where the distributions are superimposed and in the cervical area where the spines assimilate.

Like in the previous subsection, the lengths of the norm spines of the two sets can be compared using the sign test. In this case, spines of younger patients seem to be longer, i.e. straighter, than those of older patients which can be assumed due to the p-value of $1.4\text{e}{-21}$ using the appropriate hypothesis of the sign test.

The results in the coronal plane are not that definite. In many cases, the distributions only differ in their variance but seem to have a very similar mean. The KS test does not penalize this too hard which leads to p-values that do not allow to refuse the hypothesis of same distributions. One has to remember that the tested sets are not uniformly distributed but rather biased towards the mean. This way, norm spines far away from the mean do not influence the calculation as much which could cause the test to be less meaningful.

Still, the test shows significance for the x values and thus confirms the hypothesis that younger spines are straighter than older spines. Moreover, the variance of the shape seems to increase with age.

### 5.4.4 *Test on Weight*

Lastly, the influence of the weight of the patients on the norm spines will be tested. As González-Sánchez et al.[17] concluded from a dataset of 36 individuals, there is a connection between the waist circumference and the spinal curvature. To check whether this holds for the SHIP datasets as well, the hypothesis is that the distributions per height level are different.

The test that seems most suitable to check for this hypothesis is the KS test. At first, the norm spines are split up into bins with a weight range of 10 kg. Like in the previous subsections, a critical value is set first. Since the weighted mean of the weight distribution is at 80 kg, the critical value is set to this mean and so all spines with a lower weight are tested against the rest. The results can be seen in Figure 5.8.

Regarding the anteroposterior direction (x values), the result is very clear: In all cases, the hypothesis that the samples come from the same distribution can be rejected on the basis of the p-value and the common significance level of $\alpha = 0.05$. For the z values, there are some regions, i.e. the cervical and the lower lumbar area, where the hypothesis cannot be rejected. This coincides partly with the results

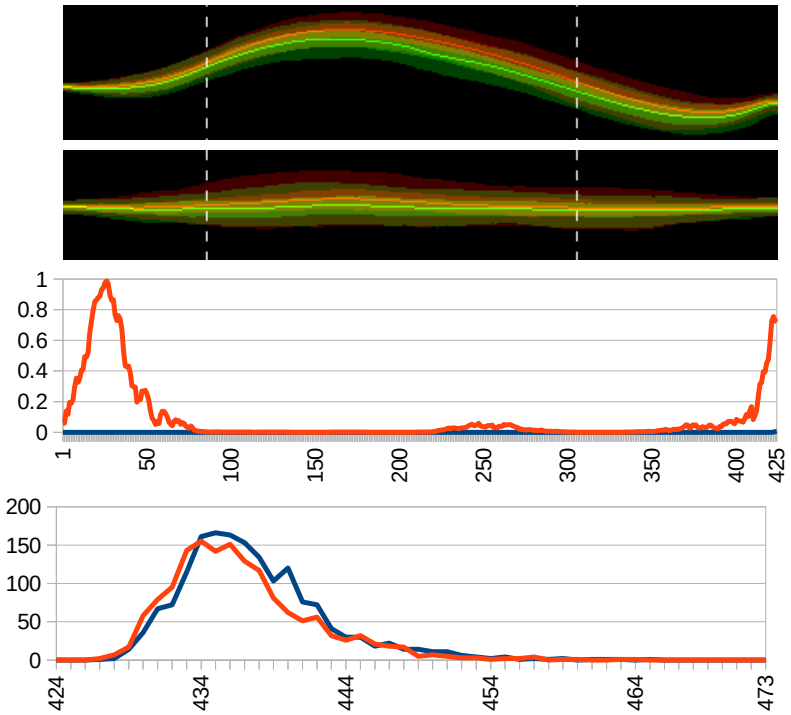


Figure 5.8: Test on Weight. Top: Sagittal and coronal projection of norm spines of bin 50 kg (red channel) and 100 kg (green channel) using 5th, 25th, 75th and 95th percentile and mean. Center: p-values of KS test for x (blue) and z (red) values per height level. Bottom: Binned distribution (1 rpx per bin) of sagittal lengths of norm spines for weight$<$75 kg (blue) and size$\geq$75 kg (red).

of González-Sánchez et al. who additionally found out that one cannot make a connection between the lateral z values and the waist circumference. In contrast to that, the KS test shows that the distributions are different in the thoracic region.

These results can also be confirmed by comparing the bins 50 kg and 100 kg visually. The sagittal projection shows that spines of 100 kg-patients are closer to zero, i.e. they are straighter. The coronal projection shows that 50 kg-spines seem to be less straight in the mid-thoracic region but also that the variance of these spines is higher compared to the 100 kg-spines.

The sign test on the lengths of the norm spines of these two sets shows that norm spines of light patients are longer, i.e. less straight, than those of heavier patients with a p-value of $2.4\mathrm{e}{-5}$ supporting this alternative. This can also be seen in the distribution of lengths in Figure 5.8 (bottom).

All in all, the performed tests support the hypothesis of unequal distributions and thus shows an influence of the weight on the curvature of the spine. Moreover, the sign test shows that heavier patients have straighter spines than light patients, on average.

## 5.5 HEALTHY SPINE CURVATURE

As described in Chapter 1, it is hard to determine a range of Cobb angles of healthy spines in the sagittal view, i.e. for the classification of kyphosis. Furthermore, it was shown in Section 5.3 that it is rather abnormal for a spine to be completely straight in the coronal view.

This leads to the question in what range the curvature of a spine can be to be classified as normal/healthy. For consistency and comparability with other scientific articles, the Cobb angles of these ranges are going to be determined and evaluated statistically.

Since the norm spines are given as a sequence of points by the proposed method, it is simple to calculate the Cobb angle. Therefore, a so-called *angle map* (see Figure 5.9), i.e. a two-dimensional symmetric square matrix of size $(\#\text{height levels} \times \#\text{height levels})$, which, at position $(i, j)$, contains the angle between the direction vectors along the norm spine at height level $i$ and $j$. The Cobb angle can then be determined by finding the maximal value on the angle map[55]. Furthermore, the position of this maximum depicts the height levels between which this angle can be found, which can be helpful for a later treatment in case of an unhealthy curvature.

### 5.5.1 *Sagittal Cobb Angles*

In the following, the ranges of angles will be evaluated per region, i.e. cervical, thoracic and lumbar area, because the spine is concave, convex and again concave in the respective regions. This will give more

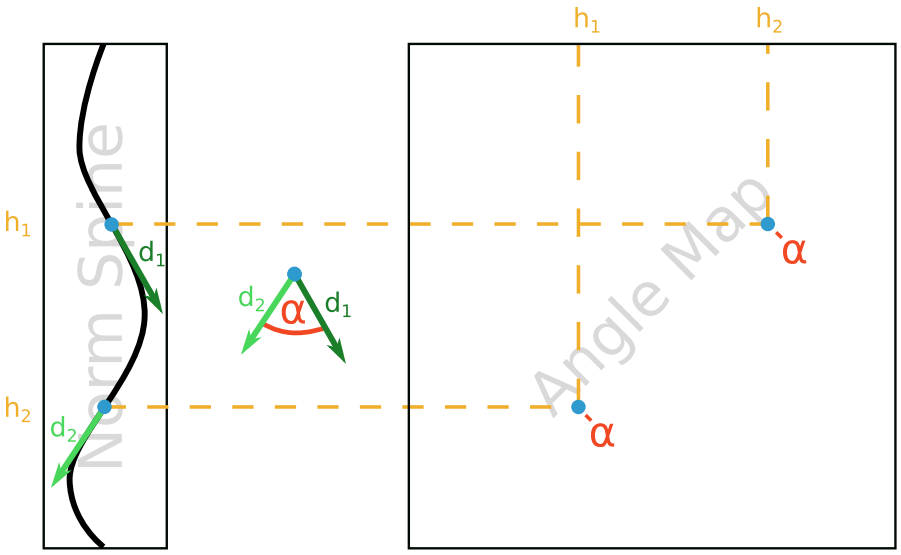


Figure 5.9: Calculation of the angle map. Left: norm spine with direction vectors along spine $d_1$ and $d_2$ at height levels $h_1$ and $h_2$. Center: Cobb angle $\alpha = \arccos(\langle d_1, d_2 \rangle / (\|d_1\| \|d_2\|))$. Right: angle map with marked Cobb angles of height levels $h_1$ and $h_2$.

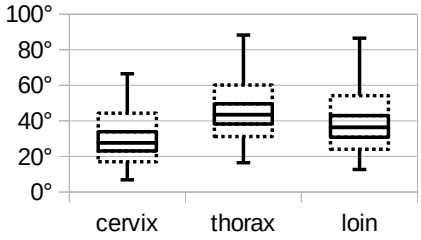


Figure 5.10: Cobb angle statistics in the sagittal view. Solid line: 1st to 3rd quantile. Dashed line: 5th percentile to 95th percentile. Error bars for minimum and maximum.

accurate and more local results. The calculations of the Cobb angles per region are very simple as well, since only the relevant submatrices of the angle map need to be considered for finding the maxima. The statistics shown in Figure 5.10 were computed by first determining the three Cobb angles per norm spine and then calculating the respective percentiles using these angles.

One can observe that the Cobb angles in the cervical and lumbar region are lower, i.e. between $17°$ and $44°$ or $24°$ and $54°$ for 90% of the spines, compared to the angles in the thoracic region between $31°$ and $60°$. Examples of the minimal, median and maximal thoracic Cobb angle in the sagittal plane can be seen in Figure 5.12.

### 5.5.2 *Coronal Cobb Angles*

In the coronal view, it is not only important to calculate the absolute value of the angles but also to include a direction indicated by the sign of the angle since the shape in this view is not necessarily either convex or concave. Therefore, at a position $(i, j)$ in the angle map, if the direction vector at height level $i$ points left, the angle between $i$ and $j$ will be positive and else negative. The resulting angle map in general is not symmetric any more due to the non-symmetric distribution of signs. The statistics on these angles can be seen in Figure 5.11.

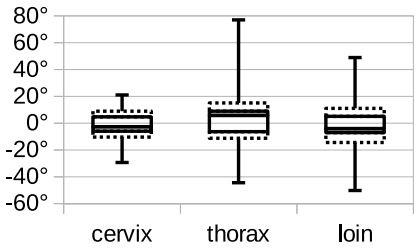


Figure 5.11: Cobb angle statistics in the coronal view. Solid line: 1st to 3rd quantile. Dashed line: 5th percentile to 95th percentile. Error bars for minimum and maximum.

Here, one can see that the angles in the three regions are always close to 0°, which means that in most cases, the spine is rather straight. For 50% of the spines, the angles in the cervical area are between -6° and 5°, for the thorax between -6° and 9° and for the lumbar region between -7° and 5°. It is also interesting to see that the medians are not very close to 0° but rather -2.5°, 5.8° and -4.1° in the respective areas. This means that it is normal to have a slightly bent spine which also supports the results of Section 5.3.

Again, examples for the minimal, median and maximal thoracic Cobb angle in the coronal plane can be taken from Figure 5.13.

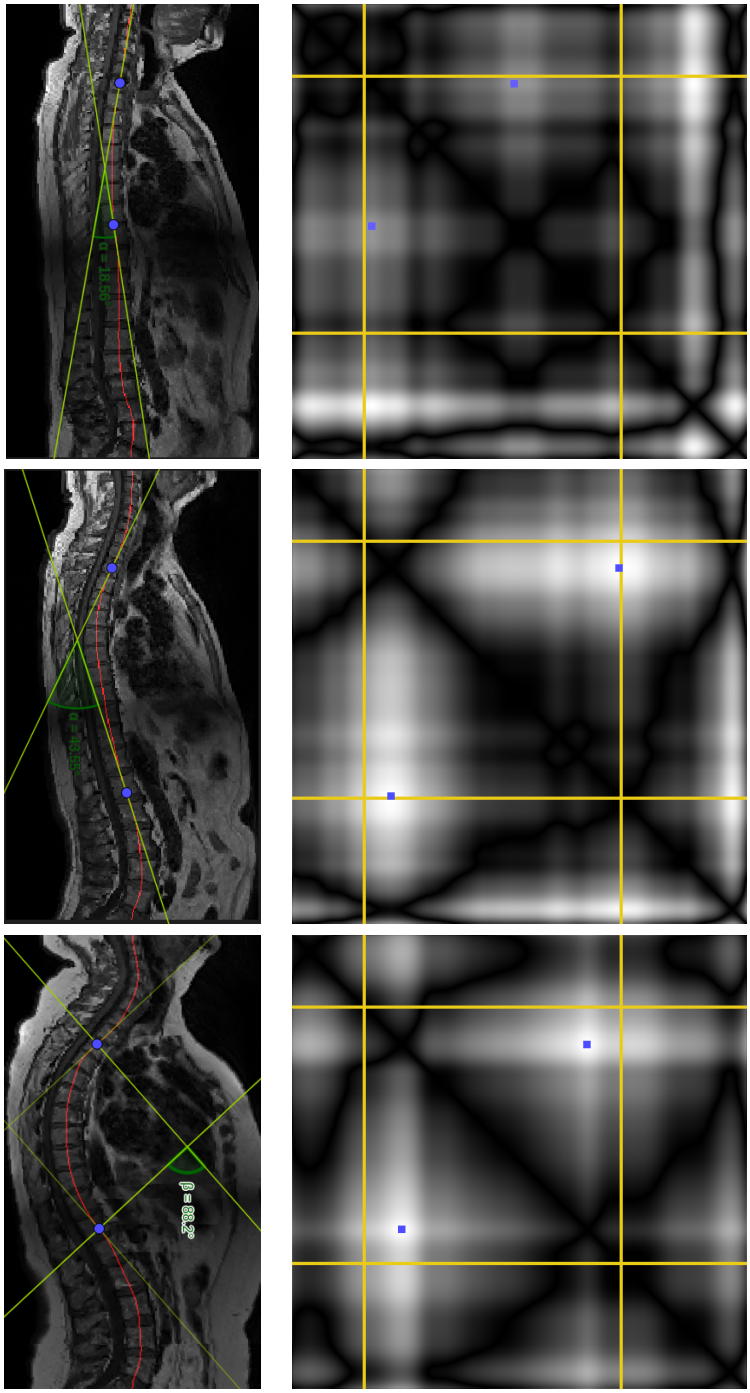


Figure 5.12: Cobb angles in sagittal view. Left column: sagittal CPR of MR record. Right column: angle map (black: $0°$, white: maximal angle in this map). Top row: minimal Cobb angle. Center row: median Cobb angle. Bottom row: maximal Cobb angle. Red: centerline. Yellow: borders between cervical, thoracic and lumbar region. Blue: points with coordinates of maxima in the angle map. Green: tangents at blue points. $\alpha = \beta$: Cobb angle in degrees.

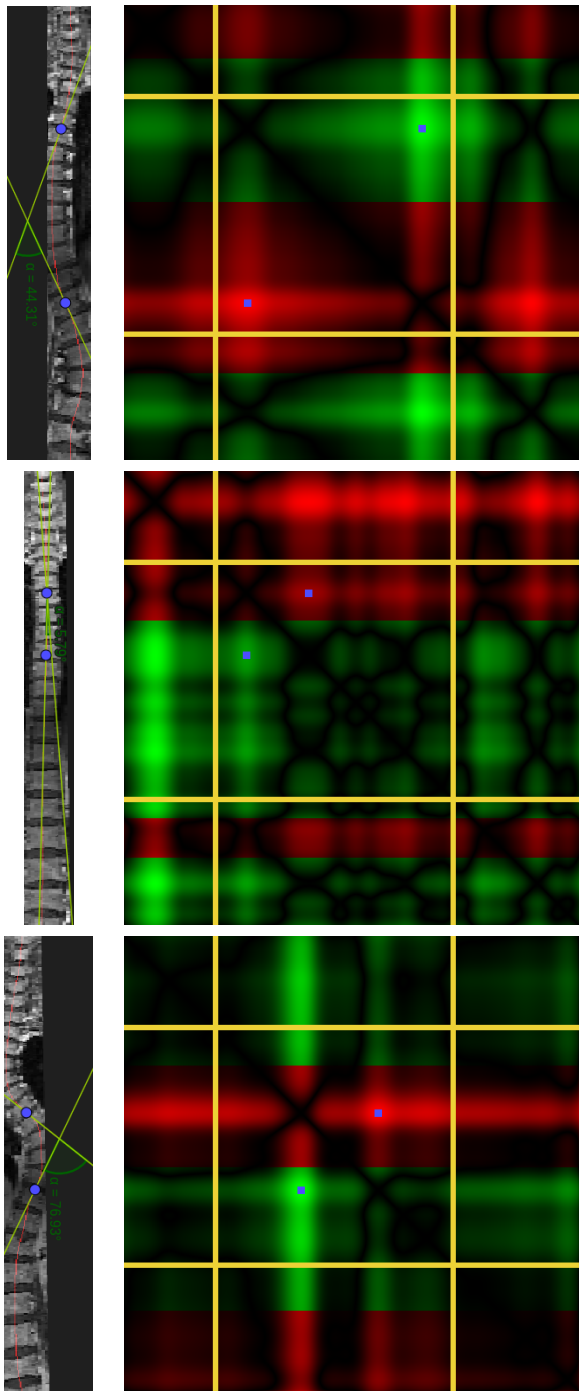

Figure 5.13: Cobb angles in coronal view. Left column: coronal CPR of MR record. Right column: angle map (green: positive angles, red: negative angles, intensity: absolute value of angles). Top row: minimal Cobb angle. Center row: median Cobb angle. Bottom row: maximal Cobb angle. Red: centerline. Yellow: borders between cervical, thoracic and lumbar region. Blue: points with coordinates of maxima/minima in the angle map. Green: tangents at blue points. $\alpha$: Cobb angle in degrees.

# 6 CONCLUSION

## 6.1 SUMMARY OF STRATEGIES

In this thesis, four strategies for a fast and sufficiently precise extraction of the the centerline of the spine from MR records using CNNs were proposed. In tests, it turned out that Strategy 2 works best for this use case, i.e. separately predicting the center of the uppermost/lowest vertebra and a so-called ridge map of the spine curve using CNNs and afterwards extracting the centerline and transforming it into the reference system.

The advantage *and* disadvantage of this strategy is the full automation once the weights of the CNNs have been trained. On the one hand, this leads to a very fast process of extracting the centerline, which makes this strategy particularly applicable on a dataset of several thousand records in a very short time. On the other hand, it prevents the user from influencing the method, e.g. if the CNN for predicting the ridge map produced an output that does not match the spinal curve in the data (see Figure 4.15).

Depending on the use case, Strategy 1, i.e. extracting a map of center points of the vertebrae using a CNN and transforming the extracted centers into the reference system, could achieve even more accurate results. Tests in Chapter 4 showed that this strategy produces slightly bigger errors in general but it would be easy to modify the method to include user interaction for manually refining the predicted centers of the vertebrae. The downside of this is of course the longer (yet still faster than manual) runtime per record since it is necessary to manually classify whether a result needs to be refined at all.

From a scientific and statistical perspective, it is feasible to accept some errors while analyzing big datasets whereas a more influenceable method would be more useful for doctors in everyday clinical routine on only a small number of records.

All in all, it could be shown in Chapter 4 that the errors of the chosen strategy for the purpose of this thesis are in most cases only 1 rpx to 1.5 rpx (i.e. 2.2 mm to 3.3 mm) for the x values and 2 rpx to 3 rpx (i.e. 2.2 mm to 3.3 mm) for the z values.

## 6.2 SUMMARY OF STATISTICAL TESTS

Statistical tests have been carried out on the extracted norm spines from two of the SHIP datasets in Chapter 5.

It could be shown that the spines in the reference system are not normally distributed per height level but nevertheless seem to have only one mode. Furthermore, the tests showed that a straight spine in the coronal view is not the optimum which means that it is not unhealthy for human beings to have a slightly bent spine, i.e. mild scoliosis, with a Cobb angle of approximately 6°. The reason for this could be due to the handedness of the population (see next section).

After the spines had been filtered for certain properties, the following results could be found. The distribution of spine shapes of male and female subjects is not the same in the coronal view. Rather, spines of female subjects seem to be straighter, visually (see Figure 5.5). It can be assumed that male spines are bent more because still, the majority of workers in typically male jobs (like mineworkers, masons and other jobs based on mainly physical work) is still male whereas most women work in mentally and rather less physically demanding jobs (like jobs in the medical sector, secretary or educator).

The body size also influences the shape of the spine: it could be shown that spines of smaller people are bent more. Of course, this directly influences each other since a person that stands straight will always be higher than when he/she bends him/herself, i.e. also bends his/her spine.

As already stated by Fon et al.[15], it could be confirmed that the spine gets bent more the older the subjects become.

Finally, it was shown that the weight of the patients has an influence on the shape as well. The tests significantly resulted in the rejection of the hypothesis that the distributions are the same per height level and even further, the sign test on the lengths of the spines showed that spines of heavier subjects are generally bent less which seems counterintuitive because one would assume that a higher mass would pull the spine down more and would thus result in a spine bent more.

Except for Fon et al.[15], no other article could be found that analyzes healthy Cobb ranges for scoliosis, kyphosis or lordosis with a statistically sufficient number of patients, which is why it was done in Section 5.5. In the sagittal view, normal ranges of Cobb angles are 20°-30°, 40°-50° and 30°-40° for cervical, thoracic and lumbar region, respectively. In the coronal view, the ranges are more or less the same for all regions, i.e. -7° to 9°.

To summarize this thesis, the goals set in Chapter 1 are evaluated:

1. Four strategies were presented to extract the curve of the spine from MR records very fast (i.e. at most 1 s per record) and precise (i.e. a maximal mean error of 3.3 mm).

2. A reference system was introduced which maps all spines onto each other. Therefore, a similarity transform is carried out on the extracted spine curves.

3. Many statistical tests on the whole dataset and on filtered subsets have been performed and showed interesting and partly unexpected results.

4. The ranges of Cobb angles for healthy spines were determined and lie between 40° and 50° in the sagittal view and -7° to 9° for the coronal view.

5. Using the angle map presented in Chapter 5, doctors and other users of this method can find out at a glance if the curvature of the examined spine is unhealthy and if so, in which area the malformation occurs.

## 6.3 FUTURE WORK

Even if detailed tests and experiments have been carried out on the strategies, there are still many spots where they could be modified and perhaps achieve even more accurate results taking almost the same time per record. Since only a coarse grid search for finding the optimal architecture of the used CNNs was performed, it is well possible to find an even better architecture, e.g. with another number and kernel size of the single convolutional layers or different activation functions.

Another point that can largely influence the accuracy of the method is the dataset, since the CNNs need to be trained. For this thesis, it was assumed that the dataset 'SHIP-Pretest' contains enough diversity and a sufficient number of records for the training process but it is possible that some larger dataset with even more variance in its data could improve the outputs to be closer to the ground truth, though this needs the creation of the ground truths at first which can be a huge effort on a big dataset.

Since only four strategies have been proposed in this thesis, it could be that an all different approach (possibly without deep learning) leads to even better results. Certainly, these methods should take a shorter time than the methods presented in Chapter 2 but yet need to be similarly accurate.

Currently, the method has only been tested on MR data. Since the CNNs are trainable, it could be possible to use this method on another kind of data (like CT or 3D ultrasound) as well. Of course, this could entail slightly changing the architecture of the networks but it is well possible to work on an appropriate dataset.

Not only could the strategies be improved but the results of the statistical tests need to be examined medically and biologically as well. For example, one question that arises is why the hypothesis of a straight coronal spine cannot be statistically confirmed. A possible reason could be that the handedness of the subjects has an effect on the spine due to an asymmetric muscle formation. Since the majority

of people in Germany is right-handed[29], the SHIP dataset cannot be seen as stratified with respect to the handedness so the median norm spine is shifted towards right-handed subjects. Nevertheless, this should be investigated in greater detail.

In Section 5.4, the data were filtered by only one property. Of course, it is possible that certain properties influence each other, e.g. body size could depend on age (older people are smaller). This means there could be interaction and moderator variables. Since this leads to $\binom{n}{2}$ possible pairs for $n$ being the number of properties, these tests have not been carried out in this thesis. Even worse, the number of mutually depending properties does not need to be 2 but could be any number which drastically increases the number of necessary tests.

Nevertheless, the tests on healthy Cobb angle ranges using angle maps can easily be carried out on a filtered subset which would, e.g., show how the Cobb angles of young people compare to those of older people.

Finally, the angle map can be used in everyday clinical practice to support doctors in determining the degree of a Cobb angle and the region of the spine where it occurs visually and very fast.

# A

# APPENDIX

## A.1 DETERMINING PARAMETERS OF THE TRANSFORM

The searched similarity transform is composed of several sub transforms that are applied one after another to the points $p_i$, $1 \leq i \leq k$, $i \in \mathbb{N}$ (sorted by height) where $k$ is the number of centers ($p_i^j$ describing point $p_i$ after applying step $j$; applying a transform to a point means in this case applying the transform to the position vector of this point):

1. Translate the points such that $p_1^1$ is the origin.
2. Scale the points with the scaling factor $s_1$ such that the line between $p_1^2$ and $p_k^2$ has a length of one.
3. Rotate the points using Rodrigues' rotation formula (resulting in the three-dimensional matrix $R$)[39] such that the line between $p_1^3$ and $p_k^3$ has the same slope as the line between $r_t$ and $r_l$.
4. Scale the points with the scaling factor $s_2$ such that the line between $p_1^4$ and $p_k^4$ has the same length as the line between $r_t$ and $r_b$.
5. Translate the points such that $p_1^5$ and $r_t$ are equal.

Since all of these transformations are affine, it is possible to combine them in one matrix:

$$A = \overbrace{\begin{bmatrix} 1 & 0 & 0 & r_{tx} \\ 0 & 1 & 0 & r_{ty} \\ 0 & 0 & 1 & r_{tz} \\ 0 & 0 & 0 & 1 \end{bmatrix}}^{\text{Step 5}} \cdot \overbrace{\begin{bmatrix} s_2 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & s_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}^{\text{Step 4}} \cdot \overbrace{\begin{bmatrix} & & & 0 \\ & R & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}^{\text{Step 3}}$$
$$\cdot \underbrace{\begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_1 & 0 & 0 \\ 0 & 0 & s_1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Step 2}} \cdot \underbrace{\begin{bmatrix} 1 & 0 & 0 & -p_{1x} \\ 0 & 1 & 0 & -p_{1y} \\ 0 & 0 & 1 & -p_{1z} \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Step 1}},$$

where $R$ needs to be determined by Rodrigues' rotation formula as stated above in Step 3.

## A.2 GRAPHS

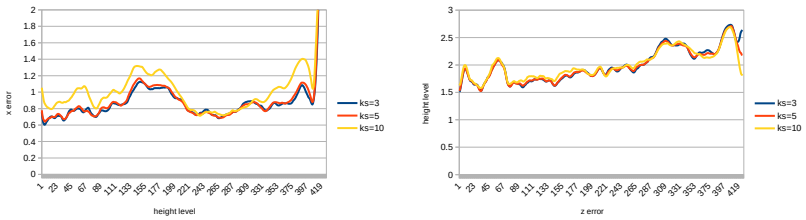


Figure A.1: Accuracy of Strategy 2. Left: Standard deviation of x values for different kernel sizes (KS). Right: Standard deviation of z values for different KSs.

## A.3 MODIFIED SIGN TEST

The standard sign test compares associated pairs of observations to check for the hypothesis for one value of a random pair to be equally likely larger than the other one.

For the groups of spine lengths, the observations cannot be associated using some common feature (like e.g. weight and size of *the same* patient) since the only feature is the length of the spine which is to be tested. Nevertheless, a random association between observations from one set to an observation of the other set can be tested (called 'modified sign test').

To verify that this random association still has the same validity as with meaningful associations, the same experiment with random associations will be executed a certain number of times. If the outcomes (i.e. the p-values) of these repetitions do not differ much, it can be assumed that the sign tests also works with random associations for these experiments.

Since the distribution of length of the norm spines visually has one mode and seems to be rather normally than uniformly distributed, it stands to reason to check the modified sign test on samples of normally distributed random variables (see Figure A.2). One can easily see that the medians of the p-values for the respective hypotheses indeed show the expected results. If the means of the two distributions are at least 0.25 apart, the modified sign test shows significance (i.e. using a significance level of $\alpha = 0.05$) for rejecting the hypothesis with the true alternative in most cases.

Since the test works on normally distributed variables, the next step is to validate this for the norm spines (see Figure A.3). One can see that the medians of p-values almost always show identical results. Nevertheless, the minimal and maximal errors can deviate a lot from the medians in some cases. One explanation for this can be the low number of samples in certain sets which do not allow a pre-

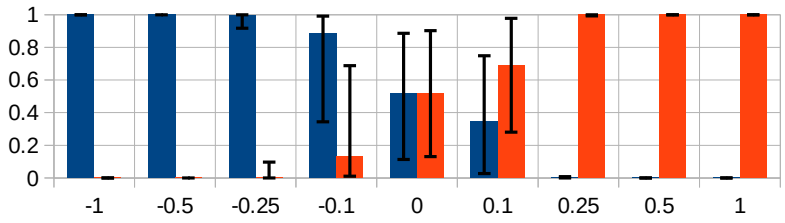


Figure A.2: Modified Sign Test on Normal Distributions: Medians of p-values after 100 shuffled repetitions for alternatives 'X>Y' (blue) and 'X<Y' (red) of normally distributed variables X and Y with the same STD of 1 using 500 samples each with error bars for minimum and maximum. The mean of Y is always 0 compared to the mean of X which is varied from -1 to 1 along the axis of abscissas.

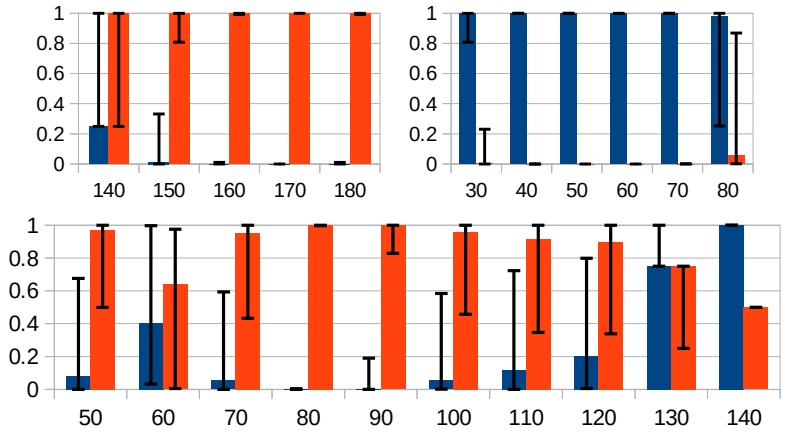


Figure A.3: Modified Sign Test on Features: Medians of p-values after 100 shuffled repetitions for alternatives 'X>Y' (blue) and 'X<Y' (red) of set X: smaller than critical value and Y: at least critical value with error bars for minimum and maximum. Critical value along axis of abscissas. Top left: Body size (in cm). Top right: Age (in years). Bottom: Weight (in kg).

cise outcome. After all, the p-values for the respective hypotheses only have a small deviation from the median for the critical values chosen in Section 5.4, i.e. 1.7 m for body size, 50 years for age and 80 kg for weight.

Because of this and because rejecting the hypothesis with some true alternative due to the p-value matches the visual results, it is possible to use the modified sign test on the lengths of the norm spines.

# BIBLIOGRAPHY


[1] H Anitha and G. K. Prabhu. "Automatic Quantification of Spinal Curvature in Scoliotic Radiograph using Image Processing." In: *Journal of Medical Systems* 36.3 (2012), pp. 1943–1951. ISSN: 1573-689X. DOI: 10.1007/s10916-011-9654-9. URL: https://doi.org/10.1007/s10916-011-9654-9.

[2] Marc A. Asher and Douglas C. Burton. "Adolescent idiopathic scoliosis: natural history and long term treatment effects." In: *Scoliosis* 1.1 (2006), p. 2. ISSN: 1748-7161. DOI: 10.1186/1748-7161-1-2. URL: https://doi.org/10.1186/1748-7161-1-2.

[3] H.P. Bischoff and F. Abdolvahab. *Praxis der konservativen Orthopädie*. Thieme, 2007, p. 425. ISBN: 9783131424617. URL: https://books.google.de/books?id=FuO0iPKmKosC.

[4] Konrad E. Bloch, Thomas Brack, and Anita K. Simonds, eds. *Self-Assessment in Respiratory Medicine*. European Respiratory Society, 2015, p. 555. ISBN: 978-1-84984-078-1. DOI: 10.1183/9781849840781-hbmcq02. URL: http://books.ersjournals.com/content/9781849840781/9781849840781.

[5] Jean-Philippe Bonjour, Gérald Theintz, Bertrand Buchs, Daniel Slosman, and René Rizzoli. "Critical Years and Stages of Puberty for Spinal and Femoral Bone Mass Accumulation during Adolescence." In: *The Journal of Clinical Endocrinology & Metabolism* 73.3 (1991), pp. 555–563. DOI: 10.1210/jcem-73-3-555. eprint: /oup/backfile/content_public/journal/jcem/73/3/10.1210_jcem-73-3-555/1/jcem0555.pdf. URL: http://dx.doi.org/10.1210/jcem-73-3-555.

[6] Norbert Boos and Max Aebi. *Spinal disorders: fundamentals of diagnosis and treatment*. Vol. 1166. Springer, 2008.

[7] Hao Chen, Chiyao Shen, Jing Qin, Dong Ni, Lin Shi, Jack C. Y. Cheng, and Pheng-Ann Heng. "Automatic Localization and Identification of Vertebrae in Spine CT via a Joint Learning Model with Deep Neural Networks." In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi. Cham: Springer International Publishing, 2015, pp. 515–522. ISBN: 978-3-319-24553-9.

[8] Dan Claudiu Ciresan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. "Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition." In: *CoRR* abs/1003.0358 (2010). arXiv: 1003.0358. URL: http://arxiv.org/abs/1003.0358.

[9] JR Cobb. "Outline for the study of scoliosis." In: *The American Academy of Orthopedic Surgeons Instructional Course Lectures*. Vol. 5. Ann Arbor, MI, 1948.

[10] S. Delorme, Y. Petit, J. A. de Guise, H. Labelle, C. E. Aubin, and J. Dansereau. "Assessment of the 3-D reconstruction and high-resolution geometrical modeling of the human skeletal trunk from 2-D radiographic images." In: *IEEE Transactions on Biomedical Engineering* 50.8 (2003), pp. 989–998. ISSN: 0018-9294. DOI: 10.1109/TBME.2003.814525.

[11] Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks." In: *Annual Conference on Medical Image Understanding and Analysis*. Springer. 2017, pp. 506–517.

[12] B. Drerup and E. Hierholzer. "Automatic localization of anatomical landmarks on the back surface and construction of a body-fixed coordinate system." In: *Journal of Biomechanics* 20.10 (1987), pp. 961 –970. ISSN: 0021-9290. DOI: https://doi.org/10.1016/0021-9290(87)90325-3. URL: http://www.sciencedirect.com/science/article/pii/0021929087903253.

[13] Luc Duong, Farida Cheriet, and Hubert Labelle. "Three-dimensional classification of spinal deformities using fuzzy clustering." In: *Spine* 31.8 (2006), pp. 923–930.

[14] Philipp Fischer, Alexey Dosovitskiy, and Thomas Brox. "Image Orientation Estimation with Convolutional Networks." In: *GCPR*. 2015.

[15] G. T. Fon, M. J. Pitt, and A. C. Thies. "Thoracic kyphosis: range in normal subjects." In: *American Journal of Roentgenology* 134.5 (1980), pp. 979–983. ISSN: 0361-803X. DOI: 10.2214/ajr.134.5.979. URL: https://doi.org/10.2214/ajr.134.5.979.

[16] Daniel Forsberg, Claes Lundström, Mats Andersson, Ludvig Vavruch, Hans Tropp, and Hans Knutsson. "Fully automatic measurements of axial vertebral rotation for assessment of spinal deformity in idiopathic scoliosis." In: *Physics in Medicine & Biology* 58.6 (2013), p. 1775. URL: http://stacks.iop.org/0031-9155/58/i=6/a=1775.

[17] Manuel González-Sánchez, Jin Luo, Raymond Lee, and Antonio I Cuesta-Vargas. "Spine curvature analysis between participants with obesity and normal weight participants: a biplanar electromagnetic device measurement." In: *BioMed research international* 2014 (2014).

[18] John A Hartigan, Pamela M Hartigan, et al. "The dip test of unimodality." In: *The annals of Statistics* 13.1 (1985), pp. 70–84.

[19] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *CoRR* abs/1502.03167 (2015). arXiv: 1502.03167. URL: http://arxiv.org/abs/1502.03167.

[20] B. Michael Kelm, Michael Wels, S. Kevin Zhou, Sascha Seifert, Michael Suehling, Yefeng Zheng, and Dorin Comaniciu. "Spine detection in CT and MR using iterated marginal space learning." In: *Medical Image Analysis* 17.8 (2013), pp. 1283 –1292. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2012.09.007. URL: http://www.sciencedirect.com/science/article/pii/S1361841512001508.

[21] Howard A King, John H Moe, David S Bradford, and Robert B Winter. "The selection of fusion levels in thoracic idiopathic scoliosis." In: *J Bone Joint Surg Am* 65.9 (1983), pp. 1302–13.

[22] Markus Rafael Konieczny, Hüsseyin Senyurt, and Rüdiger Krauspe. "Epidemiology of adolescent idiopathic scoliosis." In: *Journal of Children's Orthopaedics* 7.1 (2013), pp. 3–9. ISSN: 1863-2521. DOI: 10.1007/s11832-012-0457-4. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3566258/.

[23] S. Langensiepen, O. Semler, R. Sobottke, O. Fricke, J. Franklin, E. Schönau, and P. Eysel. "Measuring procedures to determine the Cobb angle in idiopathic scoliosis: a systematic review." In: *European Spine Journal* 22.11 (2013), pp. 2360–2371. ISSN: 1432-0932. DOI: 10.1007/s00586-013-2693-9. URL: https://doi.org/10.1007/s00586-013-2693-9.

[24] Lawrence G. Lenke, Randall R. Betz, Thomas R. Haher, Mark A. Lapp, Andrew A. Merola, Jurgen Harms, and Harry L. Shufflebarger. "Multisurgeon assessment of surgical decision-making in adolescent idiopathic scoliosis: curve classification, operative approach, and fusion levels." In: *Spine* 26 21 (2001), pp. 2347–53.

[25] Jiayun Li, Karthik V. Sarma, King Chung Ho, Arkadiusz Gertych, Beatrice S. Knudsen, and Corey W. Arnold. "A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies." In: *AMIA Annu Symp Proc* 2017 (2017), pp. 1140–1148. ISSN: 1942-597X. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977596/.

[26] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes." In: *IEEE transactions on medical imaging* (2018).

[27] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I.

Sánchez. “A Survey on Deep Learning in Medical Image Analysis.” In: *CoRR* abs/1702.05747 (2017). arXiv: 1702.05747. URL: http://arxiv.org/abs/1702.05747.

[28] R. McKenzie. *Treat Your Own Back*. Spinal Publications New Zealand, 2011. ISBN: 9780987650405. URL: https://books.google.de/books?id=m0duzgAACAAJ.

[29] Ian Christopher McManus. “The history and geography of human handedness.” In: *Language lateralization and psychosis* (2009), pp. 37–57.

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation.” In: *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE. 2016, pp. 565–571.

[31] Yurii Nesterov. “A method of solving a convex programming problem with convergence rate O(1/sqr(k)).” In: *Soviet Mathematics Doklady* 27 (1983), pp. 372–376. URL: http://www.core.ucl.ac.be/\~{}nesterov/Research/Papers/DAN83.pdf.

[32] Peter O. Newton, Takahito Fujimori, Josh Doan, Fredrick G. Reighard, Tracey P. Bastrom, and Amirhossein Misaghi. “Defining the "Three-Dimensional Sagittal Plane" in Thoracic Adolescent Idiopathic Scoliosis.” In: 97 (Oct. 2015), pp. 1694–1701.

[33] Kevin C. Parvaresh, Emily J. Osborn, Fredrick G. Reighard, Joshua Doan, Tracey P. Bastrom, and Peter O. Newton. “Predicting 3D Thoracic Kyphosis Using Traditional 2D Radiographic Measurements in Adolescent Idiopathic Scoliosis.” In: *Spine Deformity* 5.3 (2017), pp. 159 –165. ISSN: 2212-134X. DOI: https://doi.org/10.1016/j.jspd.2016.12.002. URL: http://www.sciencedirect.com/science/article/pii/S2212134X16302817.

[34] Judith MS Prewitt. “Object enhancement and extraction.” In: *Picture processing and Psychopictorics* 10.1 (1970), pp. 15–19.

[35] Ning Qian. “On the momentum term in gradient descent learning algorithms.” In: *Neural Networks* 12.1 (1999), pp. 145 –151. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(98)00116-6. URL: http://www.sciencedirect.com/science/article/pii/S0893608098001166.

[36] Miguel Reyes, Albert Clapés, José Ramírez, Juan R. Revilla, and Sergio Escalera. “Automatic digital biometry analysis based on depth maps.” In: *Computers in Industry* 64.9 (2013). Special Issue: 3D Imaging in Industry, pp. 1316 –1325. ISSN: 0166-3615. DOI: https://doi.org/10.1016/j.compind.2013.04.009. URL: http://www.sciencedirect.com/science/article/pii/S0166361513000924.

[37] Manuel D. Rigo, Mónica Villagrasa, and Dino Gallo. "A specific scoliosis classification correlating with brace treatment: description and reliability." In: *Scoliosis* 5.1 (2010), p. 1. ISSN: 1748-7161. DOI: 10.1186/1748-7161-5-1. URL: https://doi.org/10.1186/1748-7161-5-1.

[38] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method." In: *Ann. Math. Statist.* 22.3 (Sept. 1951), pp. 400–407. DOI: 10.1214/aoms/1177729586. URL: https://doi.org/10.1214/aoms/1177729586.

[39] Olinde Rodrigues. "Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace: et de la variation des cordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire." In: *Journal de mathématiques pures et appliquées*. Vol. 5. 1. 1840, pp. 380–440.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[41] Pierre Roussouly and Colin Nnadi. "Sagittal plane deformity: an overview of interpretation and management." In: *Eur Spine J* 19.11 (2010), pp. 1824–1836. ISSN: 0940-6719. DOI: 10.1007/s00586-010-1476-9. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2989270/.

[42] Tri Sardjono, Michael Wilkinson, Albert Veldhuizen, Peter Van Ooijen, Ketut E Purnama, and Gijsbertus J. Verkerke. "Automatic Cobb Angle Determination from Radiographic Images." In: Publish Ahead of Print (June 2013).

[43] Dietrich Schlenzka and Vincent Arlet. "Juvenile Kyphosis (Scheuermann's Disease)." In: *Spinal Disorders: Fundamentals of Diagnosis and Treatment*. Ed. by Norbert Boos and Max Aebi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 765–796. ISBN: 978-3-540-69091-7. DOI: 10.1007/978-3-540-69091-7_28. URL: https://doi.org/10.1007/978-3-540-69091-7_28.

[44] Isaac Jacob Schoenberg. "Contributions to the problem of approximation of equidistant data by analytic functions. Part A. On the problem of smoothing or graduation. A first class of analytic approximation formulae." In: *Quart. Appl. Math.* 4 (1946), pp. 45–99. DOI: 10.1090/qam/15914. URL: https://doi.org/10.1090/qam/15914.

[45] Samuel Sanford Shapiro and Martin B Wilk. "An analysis of variance test for normality (complete samples)." In: *Biometrika* 52.3/4 (1965), pp. 591–611.

[46] Irwin Sobel. “An Isotropic 3x3 Image Gradient Operator.” In: (Feb. 2014).

[47] William Thong, Hubert Labelle, Jesse Shen, Stefan Parent, and Samuel Kadoury. “Stacked Auto-encoders for Classification of 3D Spine Models in Adolescent Idiopathic Scoliosis.” In: *Lecture Notes in Computational Vision and Biomechanics*. Vol. 20. Sept. 2014.

[48] Elias S. Vasiliadis, Theodoros B. Grivas, and Angelos Kaspiris. “Historical overview of spinal deformities in ancient Greece.” In: *Scoliosis* 4.1 (2009), p. 6. ISSN: 1748-7161. DOI: 10.1186/1748-7161-4-6. URL: https://doi.org/10.1186/1748-7161-4-6.

[49] H. Völzke. “Study of Health in Pomerania (SHIP).” In: *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* 55.6 (2012), pp. 790–794. ISSN: 1437-1588. DOI: 10.1007/s00103-012-1483-6. URL: https://doi.org/10.1007/s00103-012-1483-6.

[50] Hongbo Wu, Chris Bailey, Parham Rasoulinejad, and Shuo Li. “Automatic Landmark Estimation for Adolescent Idiopathic Scoliosis Assessment Using BoostNet.” In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Ed. by Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne. Cham: Springer International Publishing, 2017, pp. 127–135.

[51] Dong Yang et al. “Automatic Vertebra Labeling in Large-Scale 3D CT using Deep Image-to-Image Network with Message Passing and Sparsity Regularization.” In: *CoRR* abs/1705.05998 (2017). arXiv: 1705.05998. URL: http://arxiv.org/abs/1705.05998.

[52] Jianhua Yao, Stacy O’Connor, and Ronald Summers. “Automated spinal column extraction and partitioning.” In: 2006 (May 2006), pp. 390 –393.

[53] Junhua Zhang, Edmond Lou, Lawrence H. Le, Douglas L. Hill, James V. Raso, and Yuanyuan Wang. “Automatic Cobb Measurement of Scoliosis Based on Fuzzy Hough Transform with Vertebral Shape Prior.” In: *Journal of Digital Imaging* 22.5 (2008), p. 463. ISSN: 1618-727X. DOI: 10.1007/s10278-008-9127-y. URL: https://doi.org/10.1007/s10278-008-9127-y.

[54] Junhua Zhang, Hongjian Li, Liang Lv, and Yufeng Zhang. “Computer-Aided Cobb Measurement Based on Automatic Detection of Vertebral Slopes Using Deep Neural Network.” In: *International Journal of Biomedical Imaging* 2017 (2017).

[55] Dženan Zukić, Aleš Vlasák, Jan Egger, Daniel Hořínek, Christopher Nimsky, and Andreas Kolb. “Robust Detection and Segmentation for Diagnosis of Vertebral Diseases Using Routine MR Images.” In: *Computer Graphics Forum* 33.6 (), pp. 190–204.

DOI: 10.1111/cgf.12343. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12343. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12343.