Towards Automated Age Estimation of Young Individuals

A New Computer-Based Approach Using 3D Knee MRI

Markus Auf der Mauer

Towards Automated Age Estimation of Young Individuals: A New Computer-Based Approach Using 3D Knee MRI

Vom Promotionsausschuss der Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von

Markus Auf der Mauer

aus

Caracas, Venezuela

2020

1. Gutachter: Prof. Dr. habil. Michael M. Morlock
2. Gutachter: Prof. Dr. Dennis Säring

Tag der mündlichen Prüfung: 28. Februar 2020

Berichte aus der Medizinischen Informatik und Bioinformatik

Markus Auf der Mauer

Towards Automated Age Estimation of Young Individuals

A New Computer-Based Approach Using 3D Knee MRI

Shaker Verlag
Düren 2020

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Hamburg, Techn. Univ., Diss., 2020

Printed in Germany.

ISBN 978-3-8440-7400-0
ISSN 1432-4385

Shaker Verlag GmbH • Am Langen Graben 15a • 52353 Düren
Phone: 0049/2421/99011-0 • Telefax: 0049/2421/99011-9
Internet: www.shaker.de • e-mail: info@shaker.de

Acknowledgements

> There are no secrets to success. It is the result of preparation, hard work, and learning from failure.
>
> Colin Powell

During the course of my PhD, I prepared each step of the project to be as efficient as possible, I worked hard to implement and evaluate my ideas, and I learned from failures to adapt and improve my strategies. I would like to thank the following people for making these steps possible through their guidance, support, and motivation.

First, my doctoral father Prof. Michael M. Morlock for his enthusiasm for the project and his academic guidance. I do not only appreciate his practical thinking but also the knowledge he transmitted during my studies.

My PhD supervisor Prof. Dennis Säring for his continuous, skilled, and dedicated support and guidance. His insight and knowledge in the fields of medical image processing and machine learning steered me through this research project. Our regular meetings and conversations were inspiring for me to think outside the box and pushed me in the right direction.

Eilin Jopp-van Well for her expertise in age estimation and together with my other research associates Jochen Herrmann, Michael Groth, Rainer Maas, Ben Stanczus, and Paul-Louis Pröve, for their collaborative effort and energy during data acquisition, journal publications and overall help throughout the project.

My former colleagues at the University of Applied Sciences of Wedel for their interest in my PhD work and the enriching experience working together for over three years.

Of course, my friends and family for their support, motivation, and understanding at all times. I am immensely grateful to my parents for laying the foundation to reach this milestone of my life. Nothing is more important than family.

Finally, I would like to express my deepest gratitude to my partner Julia Sabeike for always being there for me in both good and difficult moments. Your love, support, patience and understanding have made the success of this project possible.

Abstract

Background: Age estimation from medical images plays an important role in forensic medicine to determine the chronological age of individuals lacking legal documentation or to discriminate minors from adults. Current methods for imaging-based age estimation are labour-intensive, subjective, and involve radiation exposure. Recent studies indicate that magnetic resonance imaging (MRI) offers a viable alternative to established methods. The *goal of this work* is to develop a fully automated, computer-based, and non-invasive method to estimate the chronological age of male adolescents and young adults based on knee MRIs.

Materials and Methods: A total of 489 three-dimensional knee MRIs were acquired from 299 male Caucasian subjects aged 13 to 21 years. The dataset was expanded with numeric data of the subjects (anthropometric measurements and assessments of knee bone maturation). The proposed solution for automated age estimation is composed of three parts: (a) *pre-processing* to standardize the data, (b) *bone segmentation* via convolutional neural networks (CNNs) to extract age-relevant structures from the images, and (c) *age estimation*. Three different methods were investigated in part (c). *Method 1* (M1) is based on machine learning (ML) and uses the numeric data to solve the task. *Method 2* (M2) is composed of a CNN which takes in knee MRIs and outputs age predictions per image slice. Subsequently, an ML algorithm is trained on these predictions and on the numeric data to estimate a single and final age per subject. Finally, *Method 3* (M3) is a variant of M2 which incorporates the numeric data into the CNN trained on knee MRIs. Similar to M2, M3 predicts a final age per subject based on ML but using only the age predictions of the CNN.

Results: The best performing method is M2 and achieves a mean absolute error in age regression of 0.69 ± 0.47 years and an accuracy in majority classification of 90.93% using the 18-year-threshold.

Conclusions: The results demonstrate the potential of this approach for age estimation based on knee MRI and ML-techniques and is expected to improve further with the incorporation of additional datasets.

Keywords: Automated age estimation · MRI · Knee · Machine learning · Convolutional neural networks · Segmentation

Contents

List of Figures

List of Tables

List of Abbreviations

Abbrv.		Meaning
3D	=	Three-Dimensional
AE	=	Absolute Error
AGFAD	=	international and interdisciplinary study Group on Forensic Age Diagnostics of the German Society of Legal Medicine
AI	=	Artificial Intelligence
AM	=	Anthropometric Measurements
ANN	=	Artificial Neural Network
AUC	=	Area Under the Curve
BAMF	=	Bundesamt für Migration und Flüchtlinge
BFC	=	Bias Field Correction
BL	=	Baseline
BN	=	Batch Normalization
CCW	=	Counter-Clockwise
CNN	=	Convolutional Neural Network
COG	=	Center of Gravity
CPU	=	Central Processing Unit
CT	=	Computed Tomography
CV	=	Cross-Validation
CW	=	Clockwise
DCNN	=	Deep Convolutional Neural Network
DF	=	Distal Femur
DICOM	=	Digital Imaging and Communications in Medicine
DO	=	Dropout
DSC	=	Dice Similarity Coefficient
DTC	=	Decision Tree Classifier
EASO	=	European Asylum Support Office
ELU	=	Exponential Linear Unit
ETC	=	Extremely Randomized Trees Classifier
ETR	=	Extremely Randomized Trees Regressor
EU	=	European Union

Abbrv.		Meaning
FC	=	Fully-Connected
FN	=	False Negative
FNR	=	False Negative Rate
FOV	=	Field of View
FP	=	False Positive
FPR	=	False Positive Rate
FU	=	Follow-UP
GAP	=	Global Average Pooling
GBC	=	Gradient Tree Boosting Classifier
GBR	=	Gradient Tree Boosting Regressor
GIGO	=	Garbage In, Garbage Out
GMP	=	Global Max Pooling
GNB	=	Gaussian Naive Bayes
GP	=	Greulich and Pyle
GUI	=	Graphical User Interface
ID	=	Identification
IQR	=	Interquartile Range
IoU	=	Intersection-over-Union
ITK	=	Insight Segmentation and Registration Toolkit
KJC	=	Knee Joint Cavity
KNC	=	K-Nearest-Neighbors Classifier
KNN	=	K-Nearest-Neighbors
LLL	=	Lower Leg Length
LOOCV	=	Leave-One-Out Cross-Validation
LR	=	Linear Regression
LReLU	=	Leaky Rectified Linear Unit
MAE	=	Mean Absolute Error
MFS	=	Magnetic Field Strength
ML	=	Machine Learning
MLP	=	Multilayer Perceptron
MRI	=	Magnetic Resonance Imaging
MSE	=	Mean Squared Error
N3	=	Nonparametric nonuniform intensity normalization algorithm
N4ITK	=	Improved N3 algorithm for ITK

Abbrv.		Meaning
NCC	=	Normalized Cross-Correlation
OC	=	Ossification Classes
PCA	=	Principal Component Analysis
PCL	=	Posterior Cruciate Ligament
PF	=	Proximal Fibula
PReLU	=	Parametric Rectified Linear Unit
PT	=	Proximal Tibia
ReLU	=	Rectified Linear Unit
RF	=	Random Forests
RFC	=	Random Forests Classifier
RFR	=	Random Forests Regression
RMSE	=	Root Mean Squared Error
ROC	=	Receiver Operating Characteristic
ROI	=	Region of Interest
SENSE	=	SENSitivity Encoding
SGD	=	Stochastic Gradient Descent
SKJ	=	Score of the Knee Joint
SVC	=	Support-Vector Classification
SVM	=	Support-Vector Machine
SVR	=	Support-Vector Regression
TE	=	Echo Time
TN	=	True Negative
TNR	=	True Negative Rate
TP	=	True Positive
TPR	=	True Positive Rate
TR	=	Repetition Time
TSE	=	Turbo Spin Echo
TW2	=	Tanner-Whitehouse method 2
UKE	=	University Medical Center Hamburg-Eppendorf
VOI	=	Volume of Interest
VTK	=	Visualization Toolkit

1 Introduction

In recent years the European Union (EU) witnessed one of the largest migration crisis since World War II. Over 2.5 million refugees sought asylum in the EU in 2015 and 2016 alone [14], escaping civil war, poverty, or other reasons. This has posed an enormous challenge to all countries of the EU, starting from the medical and basic needs aid after long and difficult journeys up until the acceptance and intensive integration of the refugees into the target country.

The need for age estimation plays an important role in criminal proceeding and professional youth sport tournaments as well but has especially strengthened due to the European migrant crisis. Age estimation is a procedure to determine the chronological age of an individual who lacks legal documentation [12] or provides one where the authenticity is dubious [56, 62]. The most important and challenging part is the discrimination between adults and minors to protect children and provide them with the benefits they are entitled to by law [56, 62, 108, 118]. Such benefits include e.g. the accommodation in a youth facility, access to eduction, and a legal guardian. Unfortunately, there is currently no method that offers an exact determination of the chronological age of a person [56, 57, 63]. Many studies have concluded that the combination of various methods could provide an opportunity to reduce uncertainty and increase the overall reliability of the estimation [12, 49, 173]. In any case, the principle of "*in dubio pro reo*" should be followed, i.e. when in doubt of the estimated age, the decision should benefit the accused [57].

The European Asylum Support Office (EASO) has published an overview and recommendation of age estimation methods in 2013 [56] with an update in 2018 [57]. Their guideline states that the first step should be to analyze the provided legal documents and the statements by the individual. In case the evidence confirms the claimed age there is no need for age estimation. On the contrary, if documents are not available or their authenticity is doubtful, then non-medical and medical methods should be performed. Documentation is a considerable issue in the developing world as reported by Unicef since only 50% of the children under 5 years of age have their births registered [56, 212].

Non-medical methods include the analysis of further documentation related to the subject, a personal interview to deduct the chronological age from life events, and a psychological assessment. But these methods have a series of drawbacks. For example, the retrieval of additional documentation such as school or medical records might not be possible and it is unclear which type of documents should be accepted or not. The interviews and psychological assessments are labour-intensive, subjective to the professional conducting the test, and have a wide margin of error. Nevertheless, EASO suggests to perform non-medical methods before undertaking further assessments since they are physically *non-invasive*. [56, 57]

Medical methods can further be divided into methods with or without ionizing radiation. EASO and further sources state that radiation exposure to an individual should be used as a last resort since it is considered bodily harm [49, 57, 123, 168, 172].

Radiation-free medical methods include dental observation, magnetic resonance imaging (MRI) or ultrasound of various long bones of the human body, and physical development assessment. Dental observations use the third molars as an indicator of adulthood but the method can be useless if these teeth are not available at the time of examination. MRI and ultrasound are used to examine the ossification degree of the growth plates of the long bones which has shown a correlation with the chronological age in multiple studies [35, 52, 58, 91, 92, 104, 109, 143]. MRI has the disadvantage of being expensive, requiring long examination times, and not being suitable for individuals with metal in the body. EASO states that ultrasound is currently considered unsuitable for age estimation since it does not provide sufficient visualization of all bone fusion stages. Lastly, the physical development assessment compares anthropometric measurements (e.g. height and weight) and sexual maturity with reference values. It can be a traumatic procedure for individuals depending on their background and is also considered the least accurate of all forensic medical methods. [56, 57]

Medical methods using radiation include X-rays of the hand and wrist, the collar bone, and pelvic bone. These methods are also based on the ossification degree of growth plates. The third molar development can be analysed using X-ray as well. The advantages of using X-rays are that it is more accessible and economical and the examination times are generally shorter than MRI, but there is ethical opposition due to radiation since it is considered bodily harm. [56, 57]

In summary, all methods have their advantages and disadvantages. EASO states that medical methods have the advantage over non-medical ones that the margin of error is smaller and that they are considered scientifically accepted since they are based on validated studies [56, 57]. Therefore, it was decided to focus and further investigate *medical* age estimation methods. Following, are the most common drawbacks of current studies using medical methods for age estimation:

- radiation
- labour-intensive
- subjective
- based on outdated reference data
- inhomogeneous population data
- non-uniform age distribution

Radiation was already discussed and can be overcome using MRI or ultrasound. All methods, radiation-free or not, are *labour-intensive* and *subjective* since the bone maturity and physical development of the subjects have to be analyzed manually by one or multiple experts. *Old reference data* is used for most of the methods based on X-rays. For example, radiographs of the hand and wrist are compared to reference images from individuals of different age groups of the Greulich and Pyle atlas (GP) [71], which collected data from boys and girls in the 1930s [56, 67]. Similarly, another approach based on individual hand bones, the Tanner-Whitehouse method (TW2) [203] acquired data from children in the 1960s. Critics say that the maturity pace of individuals has accelerated over the last few decades due to socioeconomic factors such as nutrition, medical care, and regional differences, as well as due to environmental effects. Presently, earlier ossification of the growth plates and earlier mineralisation of the teeth are expected [17, 68, 77, 78, 102, 135, 142, 156, 174, 217]. Furthermore, many studies gather *inhomogeneous population data* in terms of sex and age range which induces higher variability and makes comparisons between studies difficult. For instance, in [35, 51, 104] the growth plates of both male and female subjects in a broad age range of 10 to 30 years of age. In [28] on the other hand, the age range is shifted to much younger subjects between 3.8 and 15.6 years. Other studies only include male subjects in a narrower age group of 14 to 20 years [92, 162]. Besides age range, different age distributions are also commonly observed and can cause comparisons to be difficult. Especially *non-uniform age distributions* can introduce a bias in the evaluation of an age estimation method.

To overcome these drawbacks, computer-based approaches can help to achieve *fully-automated* and *unbiased* analysis of medical images *on a large scale*. For example, they could be used to automatically determine the growth plate ossification degree from MR or X-ray images and then use the result for age estimation. Alternatively, one could make use of artificial intelligence (AI) to teach a system to determine the chronological age of a subject directly from the image. Recent advances in AI have led to fully automated workflows and new state of the art results in the medical field [89, 122]. A popular research area in AI are artificial neural networks (ANNs). These are known to be feature selectors, meaning that they can learn to extract information that is relevant to a specific task [187]. In other words, ANNs could potentially be used to extract information from the medical images that are relevant for age estimation on their own. Recent studies have applied AI for age estimation and achieved promising results [86, 110, 115, 119, 193, 196, 197].

Furthermore, many studies have focused on simplifying age estimation based on medical images by reducing the image to the information that is relevant for the task. That includes the automated detection of bones and growth plates and the removal of background and undesired tissue in MR images. Different approaches have been used, from the detection of anatomical landmarks [195–197], the extraction of a region or volume of interest (VOI) [148, 165, 195–197], to the detection and extraction of bones [148] or growth plates [28].

1.1 Goal of the Work

The goal of this work was to develop a fully automated, computer-based, and non-invasive method to estimate the chronological age of male adolescents and young adults based on three-dimensional (3D) knee MRIs.

The *automation* allows the method to be user-independent and to be applied on a large scale. Moreover, the *computer-based* aspect of the solution enables the reproduction and verification of the method on other data collectives. Working with images acquired with *non-invasive* MRI is crucial due to the ethical issue of radiation which is present in most of the current methods. Gathering *only male subjects* in an *age range around adulthood* creates a homogeneous underlying population and removes a large amount of variability for age estimation which is already considered an “inexact science” [61].

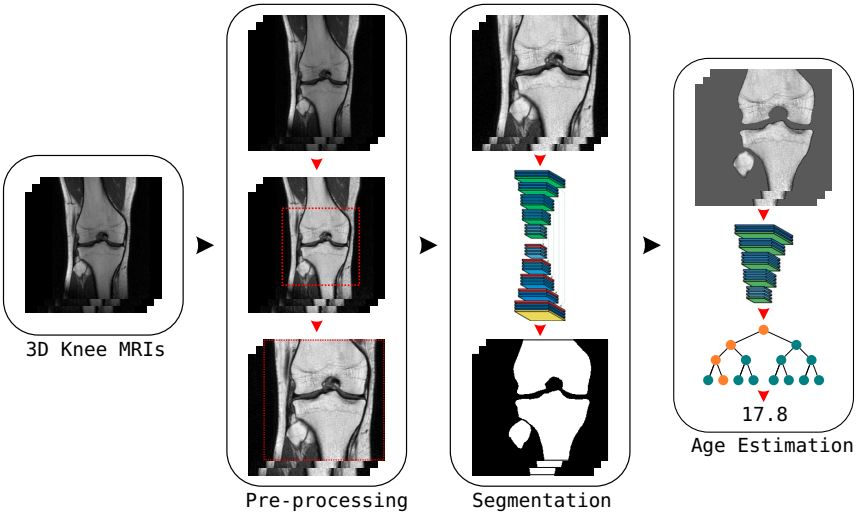

Figure 1.1: Proposed Solution for Automated Age Estimation

The proposed computer-based solution is composed of three major parts (Fig. 1.1). First, the acquired 3D knee MRIs are pre-processed to correct intensity non-uniformities frequently found in MRIs and subsequently standardized VOIs are extracted. The VOIs enable the successful extraction of age-relevant structures via bone segmentation in a second step. The motivation is to simplify the problem of estimating the chronological age of a subject from appearance information. The third and final step is age estimation based on masked images which contain only the relevant structures for the task.

1.2 Structure

At first, current age estimation methods used in practice are reviewed in chapter 2, with a focus on non-invasive and automated solutions. Next, the acquired data for this work, i.e. the knee MRIs and additional data, and the study population are described in chapter 3.

The main part of this work is the development of the methodology which is divided into three parts and follows the pipeline in Fig. 1.1. Chapter 4 provides insight to

the algorithms used and implemented to pre-process the knee MRIs. Subsequently, chapter 5 describes the approach that was developed to extract age-relevant structures from the images using convolutional neural networks (CNNs), a deep learning technique specialized for image data. The third and last methods chapter is 6 which presents multiple approaches to address the scientific research problem of estimating the chronological age of male adolescents and young adults using MRIs and other subject-related data.

Following the methods, are chapters 7 and 8 with the results and discussion for each of the main parts. Finally, the conclusions and key contributions of this work are highlighted in chapter 9. Additional information about various aspects of this work, such as available hardware, used software, and further results, can be found in Appendices A, B, C, D, and E.

2 State of the Art in Age Estimation

This chapter introduces medical methods for age estimation that are currently used in practice. The focus of this chapter lays on computer-based methods using MRI as imaging technique.

The introduction already highlighted various medical methods for age estimation. The problem in practice is that there are no standardized procedures in Europe [12, 49]. Therefore, the EASO [56, 57] and the international and interdisciplinary study Group on Forensic Age Diagnostics of the German Society of Legal Medicine (AGFAD) [123, 170, 171] have issued recommendations for criminal and asylum proceedings.

Two of the recommended methods not requiring any imaging technique are the dental and physical development examinations. Both methods have a wide margin of error of at least ± 2 years and do not take into account the variation in maturation due to nutrition, socioeconomic background, ethnicity, and more [56, 191]. Therefore, these methods are not described further.

The most common medical methods are based on the X-ray examinations of the wrist and hand using the GP or TW2 methods [71, 155, 175, 176, 203] or of the teeth (orthopantomograms) [36, 103, 125, 133]. These methods have a margin of error of 2-3 years [56, 61]. In cases of completed skeletal development of the hand and wrist and for the determination of the age limit of 21 years a computed tomography (CT) of the clavicle can be performed [173, 181, 214, 220]. However, the most important drawbacks of these methods are the old reference data and the radiation exposure. Methods based on radiation are only allowed in cases in which a judicial decision was granted [123, 168], e.g. criminal proceedings, but is less frequent for the clarification of the legal capacity of a refugee. Therefore, the EASO suggests the use of non-invasive medical methods such as MRI or ultrasound if necessary [56, 57]. Further information can be found in the aforementioned studies, which are not discussed further in this chapter.

Radiation-free methods have been analyzed in multiple studies to inspect the growth plate ossification of long bones in the body and the mineralisation of the teeth. MRI

has beed studied for age estimation based on the skeletal development of the hand and wrist [33, 46–48, 66, 180, 185, 186, 186, 205, 207, 208], of the ankle [50, 160], of the clavicle [79, 80, 177], of the third molars [6, 31–34, 72], of the spheno-occipital synchondrosis [51], and of the knee [28, 35, 52, 58, 90, 92, 104, 105, 109, 136, 143, 162, 215]. Ultrasound has also been investigated in [150, 178, 179, 182, 183]. From the publication dates, it is clear that radiation-free methods are much more recent, the oldest from 2002 in comparison to methods based on radiation which date back as far as the 1930s.

The pioneers in the age assessment by MRI of the *knee* were Jopp et al. [90, 92] and Dedouit et al. [35]. The former defined a three-stage (Fig. 2.1) and the latter a five-stage system of the knee epiphyses based on coronal T1-weighted and on coronal T2-weighted MRIs, respectively. Both studies demonstrated the correlation of bone maturation in the knee with the chronological age of living individuals. Moreover, in [35] a stage five (V) of the distal femoral and of the proximal tibial epiphysis allowed the determination of an age over 18 years for males and females. In [92], a stage three (III) of the proximal tibial epiphysis indicated individuals being 16 years or older but no assertion about the 18-year-limit could be made.

The principle of medical methods is to determine the degree of maturation of the anatomical sites by radiological assessment and then derive the chronological age of the individual from it. Generally, the "minimum-age concept" is employed and recommended [169]. The minimum age of a subject is obtained from the age minimum observed for a certain attribute value (here the degree of maturation). Consider the study by Dedouit et al. [35]: the observed ages of males with a stage four (IV) of

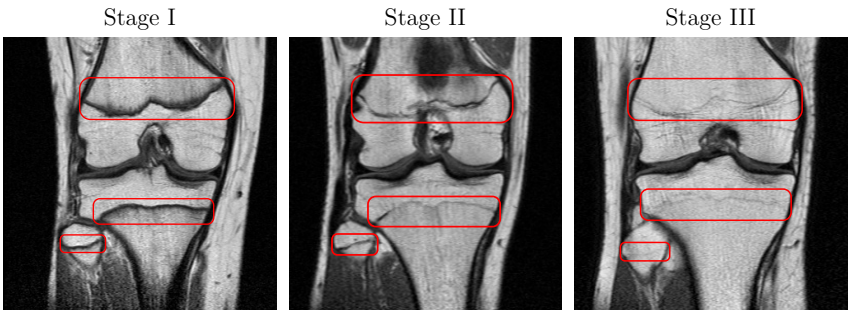

Figure 2.1: Three-stage system for the ossification degree of knee growth plates by Jopp et al. [90, 92]

the femoral epiphysis ranged between 17.8 and 30.0 years. Using the minimum-age concept, all 38 males in that study with a stage IV would be designated with an age of 17.8 years. While this concept ensures the protection of a small amount of minors, it falsely classifies a larger amount of adults as children. The oldest male in the study with stage IV was 30 years and was classified as minor with an error of 12.2 years.

Whereas the minimum-age concept of medical methods protects children from being overestimated, the difference of the estimated age to the actual chronological age can become substantially large. Additionally, all proposed methods are very labour-intensive and subjective since the anatomical sites in the images are assessed by radiologists or trained experts. To overcome these drawbacks, automated methods have been studied in the area of age estimation [161, 165, 195, 199, 213].

Saint-Martin et al. [161] analyzed the growth plate ossification in the distal tibia for 80 male and 80 female subjects (8-25 years) from T1-weighted sagittal knee MRIs. They used a method based on the analysis of gray-level variations in a previously extracted 3D region of interest (ROI) around the growth plate to discriminate between adults and minors. The authors do not specify if the ROI extraction was automated, only that it was extracted with an open source software. Their results show high specificities for both male and female subjects (above 90%) but rather low sensitivities (69% males, 62% females).

Säring et al. [165] presented an automated approach to classify the ossification degree of the proximal tibia in knee MRIs for 21 male subjects (15-19 years). The approach included the segmentation of the tibia, the extraction of a VOI around the tibial epiphysis, and the subsequent detection of large gradients. The gradients in the VOI represented the edges of the growth plates. Finally, the determination of the three ossification degree classes was performed using a Support-Vector Regression (SVR) based on 2D projection maps of gradient occurrences. Their method had an accuracy of 95.24% for the classification but no results on the application of these classes to determine the chronological age of the subjects were presented. Nevertheless, this approach could be considered as an alternative to the manual and subjective visual assessment of the growth plates performed by radiologists.

Stern et al. [195] proposed a fully automated method for age estimation based on volumetric hand MRIs from 56 male subjects (13-19 years). At first, they locate individual hand bones and their expected growth plate position in the images and

then extract a 3D ROI around the growth plate of each bone. Finally, they map intensity related features from the ROIs to the chronological age and perform a weighted sum of the estimated ages of each bone to get the final age of the subject. Their results show a mean absolute error (MAE) between the true chronological and estimated age of 0.85 ± 0.58 years. The authors state that their results are comparable to clinically established methods such as GP [71] and TW2 [203], which report values from 0.5 up to 2 years difference. In [199] they extended their dataset to 132 hand MRIs. The method was similar to the one of 2014 with the difference that they evaluated several approaches to merge the estimated ages from the individual hand bones. Their best results was a MAE of 0.82 ± 0.56 years, slightly improving their previous ones.

Further age estimation methods that are relevant to this work specialize in the AI field [86, 110, 115, 119, 193, 196–198]. In contrast to the previous chapter, methods based on deep learning can automatically learn features that are relevant to the task on their own avoiding the need to hand-craft or extract features such as the ossification degree. The works of Iglovikov et al. [86], Lee et al. [115], Larson et al. [110], and Spampinato et al. [193] all used deep learning based methods on hand radiographs. Li et al. [119] used a similar approach but on pelvic radiographs. These methods involve radiation and are mostly based on data of children and will therefore not be discussed further. Nevertheless, these studies are mentioned to show that AI is used in the age estimation field based on multiple imaging techniques.

The only studies found in literature that use deep learning for age estimation based on non-invasive techniques are the ones by Stern et al. [196–198]. In the studies from 2017 and 2018, a multi-factorial age estimation approach using MRIs of the hand, clavicle, and teeth is employed. The authors state that the inclusion of additional anatomical sites can help to extend the age estimation range up to 25 years of age. The growth plates of various bones in the human body fully ossify at different time-points and one of the last ones are in the clavicle which are also used in practice for the determination of the age limit of 21 years. In the study from 2019 they solely used 3D hand MRIs. In [196] they acquired 103 MRIs of each anatomical site from male subjects in the age range of 13 to 24 years, in [197] they increase their dataset to 322 MRIs from male subjects in the age range of 13 to 25 years, and in [198] the dataset consisted of 328 male subjects between 13 and 25 years of age. Their preprocessing of the medical images included the automated detection of the bones with a subsequent localization of the growth plates and the extraction of a 3D

ROI. In [196] and [198] they used both Random Forests (RF) [10, 13, 29] and Deep Convolutional Neural Networks (DCNNs) [23, 69, 106] for age estimation while in [197] only DCNNs. Their best results were a MAE between the true and predicted chronological age of 1.14 ± 0.96 in [196], 1.01 ± 0.74 in [197], and 0.82 ± 0.65 in [198]. In addition, they evaluated the majority classification in both studies leading to the most accurate results in terms of accuracy, sensitivity, and specificity of 91.3%, 88.6%, and 93.2% in [196] and 90.7%, 82.1%, and 96.8% in [197], respectively. In [198] majority age classification was also performed but the focus lay in determining different thresholds for the *biological age as estimated by radiologists*, not to be mistaken with the chronological age, to see the trade-off between sensitivity and specificity.

In summary, one can appreciate how methods are evolving into more reliable, accurate, and reproducible solutions for age estimation in practice. Automation is becoming the focus of recent studies as well as computer-based techniques and non-invasive imaging modalities. For a comparison to the work presented in this report, the works by Stern et al. serve as reference studies. They acquired data using MRI, evaluated multiple machine and deep learning models, and especially, targeted a vital group and age range in forensic age estimation, i.e. males between 13 and 25 years. The data and population analyzed for the current work are similar and are described in the next chapter.

3 Materials

The data for the current work includes anthropometric measurements (AM), MRIs of the knee, and growth plate ossification stages (OS) of all three knee bones determined by radiological assessment based on the MRIs. The data was acquired from three different studies and are denominated as *Dataset A* [92], *Dataset B* [3], and *Dataset C*. In total, a large data collective was available with 489 three-dimensional MRIs from 299 male Caucasian subjects with known chronological age between 13 and 21 years. AM and OS were only collected for a sub-sample of 191 individuals from *Dataset A* and *Dataset B*.

Dataset A contains data from all three categories (MRI, AM, OS) and was acquired prospectively for this work between April 2015 and June 2017 at the University Medical Center Hamburg-Eppendorf (UKE). It is a longitudinal dataset and includes data from 40 subjects for up to three time points. The time gap between each acquisition was 11 months on average (8-14 months).

Dataset B is also longitudinal and has data in all three categories from 41 subjects for two time-points each. The time gap between both acquisitions was 11.7 months on average (8-17 months).

Finally, *Dataset C* is cross-sectional and composed only of MRIs of the knee from 218 subjects. The data was collected retrospectively between December 2016 and December 2018 from a radiological unit in Hamburg (Germany).

3.1 Study Population

The following criteria were established for the selection of the study participants to address the drawbacks mentioned in the introduction and attain a homogeneous study population:

- Caucasian
- male
- middle to high socioeconomic status

- raised in Hamburg (Germany) or surroundings
- age around 18 years ± at least 3 years
- no chronic diseases or bone injuries at the growth plates of the knee

Multiple studies have shown that there has been an accelerated secular trend in human growth, maturation, and development in the last decades [17, 68, 77, 78, 102, 135, 142, 156, 174, 217]. While some studies found an impact of ethnicity on skeletal development [135, 142], most studies conclude that socioeconomic factors [68, 174], nutrition [68, 156], environmental effects such as medical care [27, 156], psychosocial environment [68], and climatic differences [135] appear to have a more decisive role. In addition, gender has a well-known influence on skeletal maturity. Several studies have confirmed that girls mature faster than boys and have a different growth pattern [27, 127, 128, 217]. Fig. 3.1 depicts the average growth rate per year for girls and boys around puberty from the study by Tanner and Davies [204].

To avoid large inter-individual variabilities due to the mentioned factors, the work only included data from *Caucasian male* subjects, with a *middle to high socioeconomic bakcground*, *parents with intermediate to high professional level*, and *raised in Hamburg (Germany) or surroundings*. The Bundesamt für Migration und

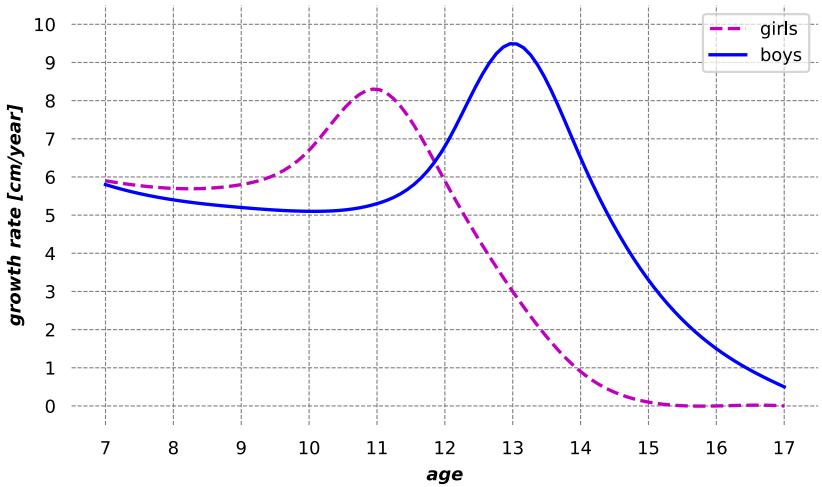

Figure 3.1: Average growth rates of boys and girls around puberty. Data from Tanner et al. [204] and Abbassi et al. [1].

Flüchtlinge (BAMF) reported that 65.7% of all asylum seekers in 2016 in Germany were male [14] which is a further reason to include only *male* subjects.

Another criteria for selection was to only include subjects which had no chronic diseases or injuries of the growth plates of the knee. Generally it can be assumed that any injury of the growth plates can result in a risk of growth disturbance [164, 184].

Multiple studies have shown that the full ossification of growth plates in the knee is possibly a suitable indicator for adulthood in males [16, 65, 104, 162] or at least for the completion of the 16th year of life [92, 105, 143]. Both are relevant legal age limits for criminal liability [16, 108]. Therefore, the objective was to target both age limits with a margin above the current state of the art ±2 years error in age estimation. The final age range of the subjects of this work was between 13 and 21 years of age. This is also a representative age range for potential asylum seekers. BAMF reported that 36.7% of asylum applicants in 2016 in Germany were between 11 and 25 years of age [14].

The age distribution can be appreciated separately for each of the three studies in Fig. 3.2. The plots of *Dataset A* and *Dataset B* are further partitioned in time-points: baseline (BL) being the first, follow-up (FU1) the second, and follow-up two (FU2) the third acquisition point per subject in time. The stacked histograms show the number of subjects per age group, e.g. "age group 15" includes subjects with a known chronological age between 15.00 and 15.99 years.

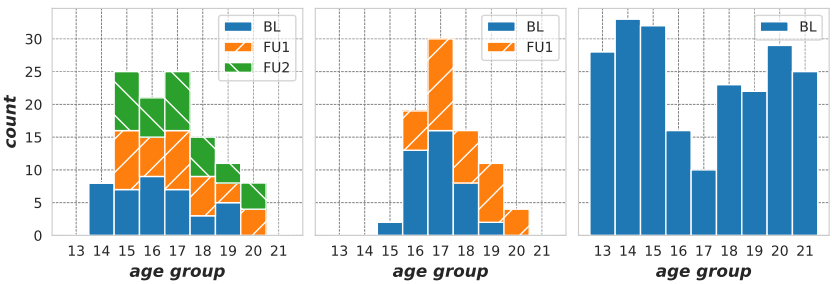

Figure 3.2: Stacked age distribution of subjects from *Dataset A* (left), *Dataset B* (middle), and *Dataset C* (left). BL, FU1, and FU2 represent different time-points for the longitudinal datasets.

3.2 Anthropometric Measurements

For the subjects *Dataset A* and *Dataset B* ($n_{A,B} = 191$), the following AM were collected: weight, standing and sitting height, and lower leg length (LLL). The weight was measured with a standard body scale device. The standing and sitting height were acquired with an anthropometer corresponding to the standardised measuring length [101, 129]. Finally, the LLL was measured using an anthropometric device [78, 93]. Fig. 3.3 depicts the sitting height and LLL. The data was retrieved at multiple time-points for each subject. Table 3.1 summarizes measurements for all subjects and time-points.

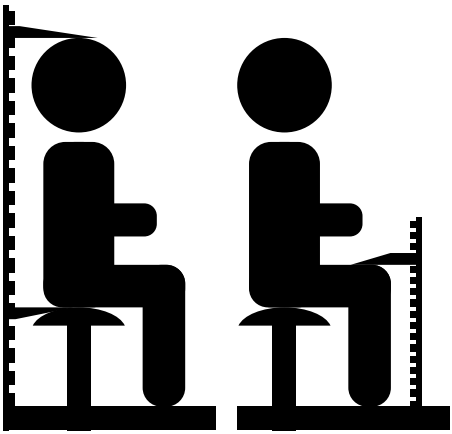

Figure 3.3: Sitting height (left), and lower leg length (right) were gathered as part of the anthropometric measurements besides weight and standing height

Table 3.1: Anthropometric measurements for male subjects of *Dataset A* and *Dataset B* aggregated for multiple time-points (age range 14-21 years)

Measurement	$n[A,B]$	Min	Max	Mean ± SD
Weight [kg]	191	48.90	124.60	71.43 ± 14.67
Standing height [cm]	191	154.50	196.00	178.66 ± 7.88
Sitting height [cm]	191	77.20	103.00	93.14 ± 4.23
LLL [mm]	191	489.50	625.50	556.88 ± 30.24

3.3 Knee MRIs

The MRI database is composed of 185 three-dimensional, coronal, and T1-weighted and 404 three-dimensional, sagittal, and T1-weighted MRIs of the knee (Fig. 3.4).

The MR images were acquired in the international standard data format for medical images known as *Digital Imaging and Communications in Medicine* (DICOM®) [38]. A DICOM data object contains several attributes, e.g. patient information (name, age, sex, etc.), acquisition device, image modality, study information, the image pixel data itself, and more. In order to comply with the data protection law, all patient related information was removed from the DICOM files prior to the import of the data from the compact discs to the workstation. To associate the files with the study participants, a number was randomly assigned to each one before the acquisition. The patient identification (ID) was set in the corresponding DICOM data object.

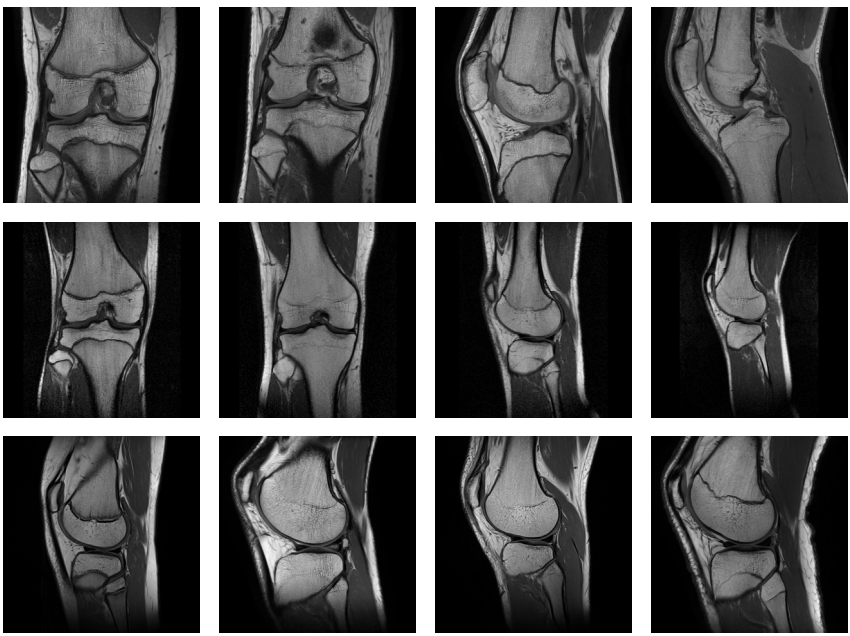

Figure 3.4: MRI slices of *Dataset A* (top row), *Dataset B* (middle row), and *Dataset C* (bottom row)

Table 3.2: Overview of the study population, the three datasets, and the MRI sequences available for this work.

	Dataset A [3]	*Dataset B* [92]	*Dataset C*
Study type	Longitudinal	Longitudinal	Cross-sectional
Subjects	40	41	218
Age range [y]	14 - 21	15 - 20	13 - 21
Time-points	0-3	2	1
Time gap [m]	8 - 14	8 - 17	-
Manufacturer	Philips	Philips	Philips, Siemens
Models	Ingenia 3.0T	Intera 1.5T Intera 3.0T	Intera 1.5T Ingenia 3.0T, Skyra 3.0T
Orientation	Coronal	Coronal	Coronal
MRIs	103	82	0
Resolution [vox.]	$800 \times 800 \times 41$	$512 \times 512 \times 24$	-
Voxel Size [mm^3]	$0.19 \times 0.19 \times 2.2$	$0.39 \times 0.39 \times 3.3$ - $0.49 \times 0.49 \times 4.9$	-
Orientation	Sagittal	Sagittal	Sagittal
MRIs	104	82	218
Resolution [vox.]	$864 \times 864 \times 50$	$512 \times 512 \times 24$	$512 \times 512 \times 24$ - $1050 \times 1050 \times 32$
Voxel Size [mm^3]	$0.17 \times 0.17 \times 2.2$	$0.39 \times 0.39 \times 3.3$ - $0.49 \times 0.49 \times 5.2$	$0.18 \times 0.18 \times 2.75$ - $0.39 \times 0.39 \times 4.38$

y := years, m := months, vox. := voxels

MRIs from *Dataset A* were retrieved with the same 3T-MRI-scanner (Ingenia 3.0, Philips Medical Systems, Best, Netherlands) and knee coil (8-Channel-Knee-Coil, Philips Medical Systems, Best, Netherlands) at all three time-points. The protocols included a T1-weighted SENSE (SENSitivity Encoding) sequence in coronal orientation (TR 850 ms; TE 10 ms; flip angle 90°) and in sagittal orientation (TR 1118 ms; TE 10 ms; flip angle 90°).

MRIs from *Dataset B* were acquired with two MRI devices. First, a 1.5T-MR-scanner (Intera 1.5, Philips Medical Systems, Best, Netherlands) and included a T1-weighted Turbo Spin Echo (TSE) sequence in coronal and sagittal orientation (TR 650 ms; TE 15 ms; flip angle 90°). Second, a 3.0T-MR-scanner (Intera 3.0, Philips Medical Systems, Best, Netherlands) and included a T1-weighted TSE se-

quence in coronal and orientation (TR 604 ms; TE 19 ms; flip angle 90°) and in sagittal orientation (TR 588 ms; TE 20 ms; flip angle 90°).

MRIs from *Dataset C* were retrieved with three MR-scanners. First, a 1.5T-MR-scanner (Intera 1.5, Philips Medical Systems, Best, Netherlands) was used for a T1-weighted TSE sequence in sagittal orientation (TR 600 ms; TE 15 ms; flip angle 90°). The specific model of the knee coil was not specified in the DICOM data. Second, a 3.0T-MR-scanner (Ingenia 3.0, Philips Medical Systems, Best, Netherlands) with a knee coil (16-Channel-Knee-Coil, Philips Medical Systems, Best, Netherlands). The protocol for the second scanner was a T1-weighted SENSE sequence in sagittal orientation (TR 1800 ms; TE 30 ms; flip angle 90°). Third, a 3.0T-MR-scanner (Skyra 3.0, Siemens, Erlangen, Germany) with a knee coil (15-Channel-Knee-Coil, Siemens, Erlangen, Germany). The protocol of the last scanner was a T1-weighted TSE sequence in sagittal orientation (TR 3500-5000 ms; TE 30-70 ms; flip angle 125-150°).

Refer to Table 3.2 for more details on the complete MRI database including the amount of MRIs per dataset and further information on the images.

3.4 Growth Plate Ossification Stages

The ossification stage of the growth plates (i.e. epiphyses) of the distal femur (DF), proximal tibia (PT), and proximal fibula (PF) was qualitatively analyzed for the coronal MRIs of sources *A* and *B*. Three raters, with over five years experience in reading MR images of the knee, independently and blindly inspected each coronal MRI in random order. The visual inspection was performed on preselected central 2D images. A fourth member, also with more than five years experience in the field, performed the selection. For both the femur and tibia, one slice was defined. The fibula was rated based on a different slice to capture the center of the bone.

The three-stage system by Jopp et al. [90, 92] was used (Fig. 3.5). This staging system fitted the T1-weighted MRI sequence in coronal slice orientation and was applied for all three epiphyses (DF, PT, PF):

- *Stage I:* epiphysis not fused
- *Stage II:* epiphysis partially fused, and epiphyseal scar is partially visible
- *Stage III:* epiphysis fully ossified, and traces of epiphyseal scar may be visible

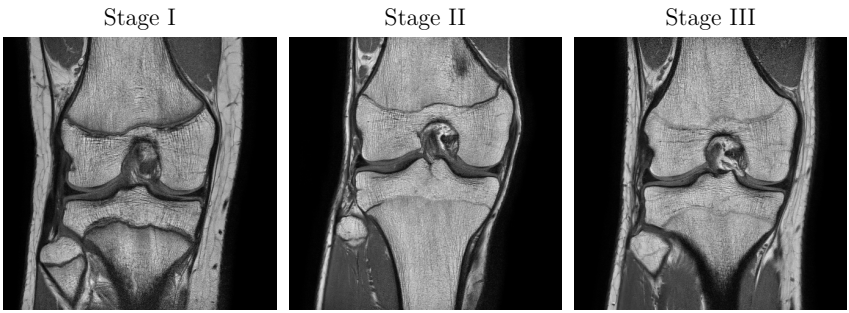

Figure 3.5: The three-stage system by Jopp et al. [90, 92] was used to assess the growth plate ossification degree in the knee and was applied to all epiphyses. Stage I (left) shows completely open, Stage II (middle) centrally ossifying, and Stage III (right) completely fused epiphyses.

Finally, the median stage of the three observers was assigned as stage to each MRI. Moreover, motivated by the work of Galic et al. [65] and Cameriere et al. [16], the classifications from all three bones were merged to gain the overall score of the knee joint, defined by the mentioned authors as the *SKJ*.

4 Image Pre-Processing

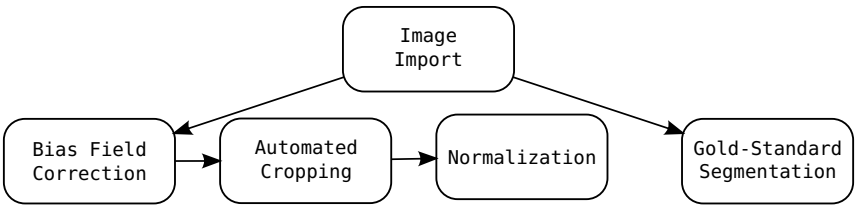

Figure 4.1: Image pre-processing for 3D knee MRIs

This chapter presents the *image pre-processing* techniques that were implemented and applied to the acquired knee MRIs. The motivation of performing pre-processing of data comes from a concept named "garbage in, garbage out" (GIGO) first mentioned in 1957 [130]. The expression means that if inaccurate or erroneous data is supplied to a computer program ("garbage in") it can lead to wrong and possibly misleading results ("garbage out"). In other words, the quality of a method can generally only be as good as the quality of the input data. Especially for supervised learning algorithms this effect can be crucial since they draw direct conclusions from the data. However, the concept cannot be applied universally as it depends on the underlying problem. For example, a system could extract or learn important information from inaccurate data to be resilient to it or detect it in new data.

Five pre-processing modules were employed in the current work (Fig. 4.1). The first module is the *image import* (section 4.1) which imports the medical image files from the data storage device to the workstation and then converts them to a file format supported by the upcoming pre-processing steps. The second module is the *bias field correction* (section 4.2) and is responsible for the correction of intensity non-uniformities in MR images. The third module is the *automated cropping* (section 4.3), responsible for the automated detection and extraction of a VOI around the growth plates of the knee. The fourth image pre-processing module is the *normalization* (section 4.4), which is a method to standardize image voxel intensities acquired from different MR-scanners. The fifth and last module is the *gold-standard segmentation* (section 4.5) which is the manual labelling procedure of bone structures performed for the 3D knee MR images.

4.1 Image Import

There are several tools available to import medical image data in DICOM format, e.g. MeVisLab [131] and 3D Slicer [15, 60]. While these tools have all the necessary functionalities to import DICOM data, they were not suitable to connect to the workflow of this work. Therefore, a tool to import DICOM data was developed and used available libraries from the *Insight Segmentation and Registration Toolkit*[1] (ITK) and the *Visualization Toolkit*[2] (VTK). These are open-source software systems for image processing and analysis and are very popular for medical applications.

The main purpose of the developed tool was to import DICOM data objects and to convert them to MetaImages. A MetaImage is a special medicine image format composed of a header file (*.mhd*) and the uncompressed image data (*.raw*). The header file contains information about the image properties such as size and spacing (Fig. 4.2). This file format was selected for further image processing since it is supported by ITK and VTK libraries.

```
ObjectType = Image
NDims = 3
ElementSpacing = 0.390625 0.390625 3.2999999999999998
DimSize = 512 512 24
ElementType = MET_SHORT
ElementDataFile = image.raw
```

Figure 4.2: Contents of a MetaImage header file (*.mhd*)

The knee image data of this work was acquired using different MRI sequences (weighting and orientation) and produced multiple cross-sectional images for each sequence (chapter 3). Thus, the DICOM images had to be correctly associated with the corresponding sequence. The tool identifies an MRI sequence, defined as a series in DICOM, by it's ID and then finds and sorts all associated files in the same folder using an ITK class[3].

Each series is then read[4] and the MRI sequence weighting and slice orientation are automatically determined from DICOM attributes. All files associated to a series

[1] https://itk.org/
[2] https://vtk.org/
[3] https://itk.org/Doxygen49/html/classitk_1_1GDCMSeriesFileNames.html
[4] https://itk.org/Doxygen49/html/classitk_1_1GDCMImageIO.html

are then read[5], converted to an internal image data format[6], and then added to a vector of 2D images. The final two steps are to combine the image slices of a series to a 3D image and to save it as a MetaImage. The pseudo-code for the implemented tool is shown in Algorithm 1.

Algorithm 1: Import DICOM data and save as MetaImage

```
Input: Folder with DICOM images of the knee
Output: MetaImage(s) (.mhd/.raw)
find all images associated to a DICOM series
for each series do
    read series
    save DICOM attributes
    determine MRI sequence weighting
    determine MRI slice orientation
    initialize series vector
    for each cross-sectional slice of series do
        read image slice
        convert to internal image format
        add to series vector
    end
    convert series vector to 3D image
    save as MetaImage with a template filename for different uses
end
```

An important feature of the tool is that it saves the MetaImages to the hard disc with a template filename (Fig. 4.3). The idea is to give the user all the necessary information about the subject and the image data without the need to open the image. Additionally, subsequent image processing algorithms can easily extract information out of the filename and apply specific rules depending if the image is e.g. in sagittal or coronal orientation. Furthermore, the template filename can help to build a database in the future which would ease the access and extraction of specific sub-samples of the data. For example, one could locate all subjects of a certain age or find all images of a specific MRI sequence.

[5] `https://itk.org/Doxygen49/html/classitk_1_1ImageFileReader.html`
[6] `https://vtk.org/doc/release/7.0/html/classvtkImageData.html`

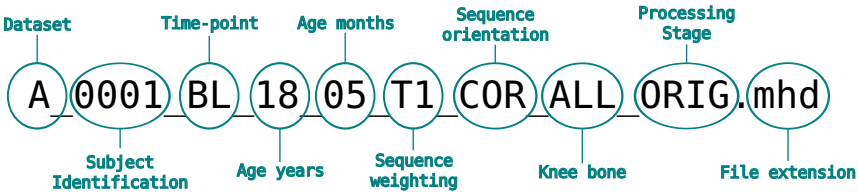

Figure 4.3: An example for a template filename of a MetaImage generated with the implemented DICOM importer

4.2 Bias Field Correction

The second image pre-processing module is *Bias Field Correction* (BFC) which corrects intensity non-uniformities in the imported 3D knee MRIs of this work (Fig. 4.1). *Intensity non-uniformity* or *bias field* is a smooth low-frequency signal that degrades MR images. It is not only caused by various coils of the MR-scanner, but also by the choice of the MR sequence and the geometry and movement of the patient [7, 84, 216]. The "corrupted" images frequently exhibit blurs and differences in intensity values of voxels of the same tissue [84, 94, 216]. While this is generally not an issue for clinical diagnosis, it can greatly decrease the performance of image processing algorithms [94, 216].

Many methods exist to correct bias fields and are mostly based on *lowpass filtering*, *surface fitting*, *statistical modelling*, *histogram*, and *segmentation* [84, 216]. A large overview can be found in the reviews by Belaroussi et al. [7], Hou [84], and Vovk, Pernus, and Likar [216]. The de facto standard in the field is N4ITK [211], an improved version of the *nonparametric nonuniform intensity normalization (N3)* algorithm [190]. The principal idea of N4ITK is to estimate the bias field present in the image and then use it to correct the corrupted image. The method assumes a noise-free scenario and thus MR artefacts that are not related to magnetic field inhomogeneities, such as patient motion, cannot be corrected.

N4ITK proved to be effective for the BFC of the 3D knee MRIs of this work (Fig. 4.4). It was implemented with the help of an ITK class[7] which is based on the paper by Tustison et al. [211]. The class is computationally expensive and scales with image size. Therefore, the images are first downsampled to a size of $448 \times 448 \times z$, where

[7] `https://itk.org/Doxygen49/html/classitk_1_1N4BiasFieldCorrectionImageFilter.html`

z stands for the number of slices of the corresponding 3D image. Then, foreground is separated[8] from background to further reduce the computational cost. Finally, N4ITK uses the foreground to estimate the bias field in the original image and subsequently corrects it. The pseudo-code for the implementation of the BFC can be found in Algorithm 2.

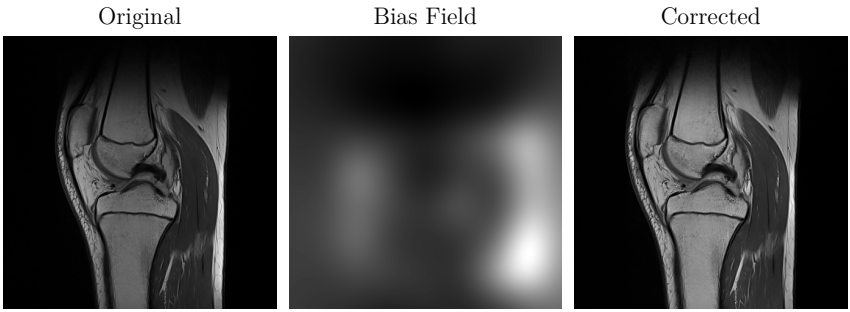

Figure 4.4: Bias field correction of a knee MR image slice. The corrected image (right) shows improved contrast in comparison to the original image (left). The estimated bias field (center) shows an over-exposed region to the left and an under-exposed area to the top of the image slice.

Algorithm 2: Bias Field Correction

Input: MetaImage (.mhd/.raw)
Output: uncorrupted MetaImage
downsample each image slice to 448×448 pixels
separate image foreground and background
execute N4ITK algorithm to correct bias field in input MetaImage

4.3 Automated Cropping

The third data preprocessing module is the automated cropping (Fig. 4.1). It generates standardized VOIs around the growth plates of the knee regardless of the field of view (FOV) of the original MRIs (Fig. 4.5). The cropping is useful to reduce the amount of undesired anatomical structures in the image (e.g. fat and muscles) and

[8] `https://itk.org/Doxygen49/html/classitk_1_1OtsuThresholdImageFilter.html`

to increase the size ratio of the target structures, i.e. the growth plates of the knee, with respect to the entire image.

To find the exact location of the VOI in the image an approach named *Patch Matching*, also *Template Matching*, was employed. The principle of this approach is to determine correspondences between a characteristic region, i.e. a *patch* or *template*, and equally sized regions across the entire target image [5, 11]. The optimal location of the patch in the image is then determined at the point of it's highest correlation with the image. The *normalized cross-correlation* (NCC) is used as a similarity measure since it is invariant to local changes in brightness and contrast [11, 228].

To enable the automated cropping of this work, characteristic regions were selected in the knee MRIs. The *intercondyloid eminence* was defined as patch for MRIs in coronal orientation (Fig. 4.6). For sagittal MRIs, the *posterior cruciate ligament* (PCL) was defined as characteristic region (Fig 4.7).

The patches were extracted from downsampled and uncorrupted images resulting from the previous pre-processing modules. Moreover, two patches for each image

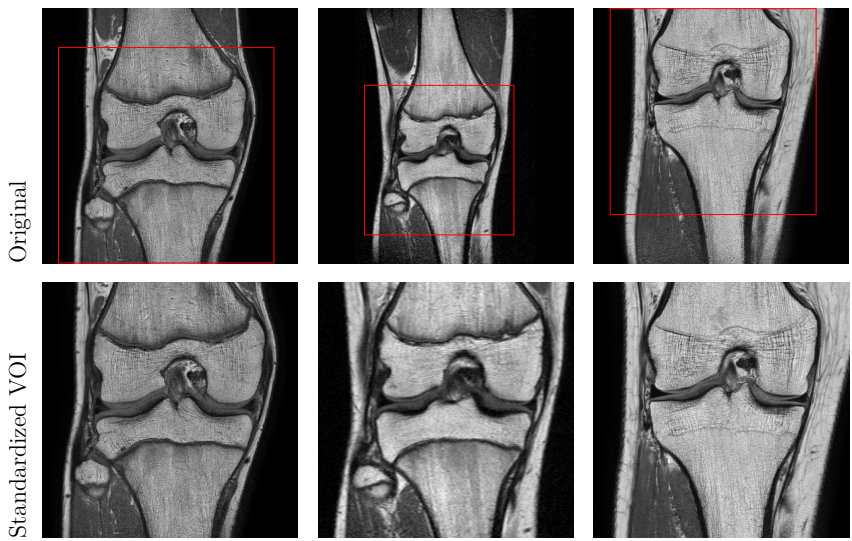

Figure 4.5: Automated cropping is used to generate standardized volumes of interest (bottom row) around the knee growth plates regardless of the FOVs of the original MRIs (top row)

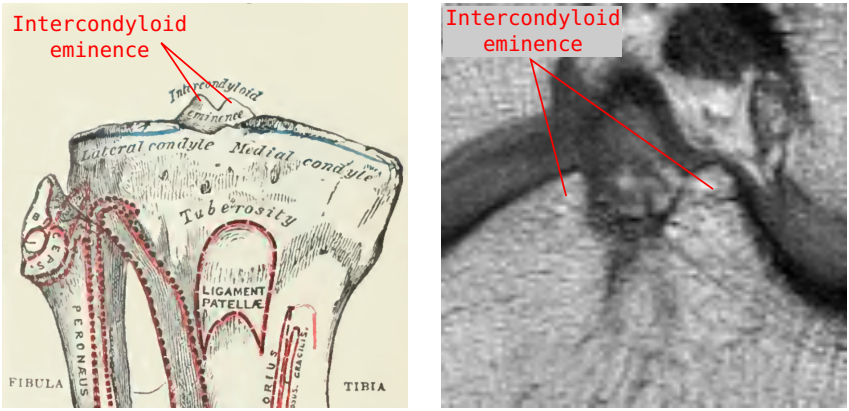

Figure 4.6: *Intercondyloid eminence* of the proximal tibia (left; adapted from [70]) and it's representation in a coronal knee MRI slice (right)

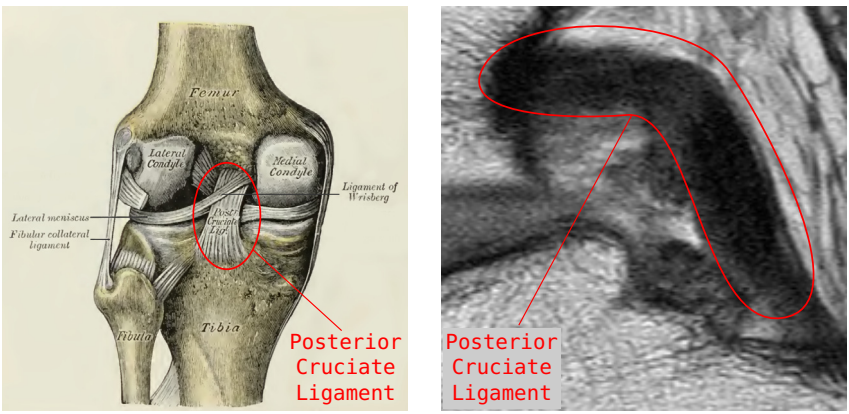

Figure 4.7: Posterior view of the knee joint with the *posterior cruciate ligament* highlighted (left; adapted from [70]) and its representation in a sagittal knee MRI slice (right)

orientation were extracted to account for spacing and resolution differences between images of the three datasets (Table 3.2). One for lower in-plane resolutions at around 0.4×0.4 mm^2 and another one for higher resolutions at approximately 0.2×0.2 mm^2. The spacing encoded in the MetaImage and the template filename enabled the automatic selection of the correct patch for a given image.

The implementation of the automated cropping is composed of multiple parts (Fig. 4.8). First, a volume smaller in size than the MRIs is defined to search for the best location of the patch in the image. This reduces the computational cost of the sliding-window principle of patch matching and increases it's robustness. In the z-dimension, i.e. along the slicing direction, 12 central slices was sufficient to successfully detect the intercondyloid eminence and the PCL in coronal and sagittal MRIs, respectively. In the x-y-plane, the search area covered 75% of the image slice (green area in Fig. 4.8). This setting covered the variations of the positions of the characteristic anatomical structures in all MRIs.

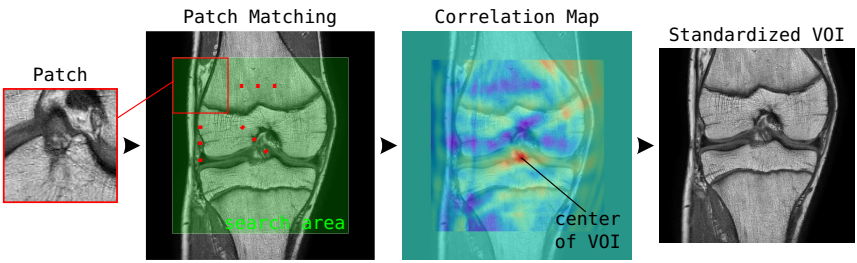

Figure 4.8: *Patch Matching* between a small characteristic region or *patch* (left) and equally sized image regions (center-left). The patch is slid across the search area (green) and for each pixel the normalized cross-correlation between the patch and the image region is computed. The brightest point in the correlation map (center-right) represents the best location of the patch in the image. A standardized VOI is built around this point (right).

The second part of automated cropping is to slide the patch across the search volume and compute the NCC at each position.[9] Afterwards, the maximum value NCC_{max} per slice z and it's position (x,y) is determined and saved to a vector $\mathbf{v}$:

$$\mathbf{v}(z) = \begin{bmatrix} x & y & \mathrm{NCC}_{max} \end{bmatrix} \tag{4.1}$$

Choosing the position with the highest overall NCC among all twelve slices, led to a false localization of the patch in a number of cases. Therefore, a refined selection of the best position was necessary. In an iterative approach, three adjacent slices of the search volume are selected and their average NCC is calculated. The three adjacent slices with the highest average NCC are shortlisted. Then, the best vertical position y_{best} is calculated as the median of the y's of the three shortlisted vectors (equation 4.1). The median of the y's was chosen here over the mean, since it was more robust to erroneous localizations. To finalize, x_{best} is extracted from the vector containing y_{best}.

The third and final step is to crop the image.[10] For this purpose, a VOI is formed with it's center at (x_{best},y_{best},z_{best}). The size of the VOI was defined in mm and not in pixels, since the images from the three datasets (A, B, C) had different spatial resolutions. The final size of the VOI was $130 \times 130 \times \mathrm{d}_z$ mm^3. The VOI included the entire image volume depth d_z, i.e. all slices, indicating that the images were in fact only cropped in-plane. The pseudo-code for the automated cropping can be found in Algorithm 3.

[9] `https://itk.org/Doxygen49/html/classitk_1_1NormalizedCorrelationImageFilter.html`

[10] When preparing the data for the training of the segmentation network, the cropping was performed after the image augmentation instead (see section 5.2)

Algorithm 3: Automated cropping

Input: MetaImage and Patch
Output: Cropped Image
define search area
for *each image slice* **do**
  for *each image position* **do**
    compute normalized cross-correlation (NCC)
  end
  determine the maximum NCC per slice z
  save to vector $\mathbf{v}(z) = \begin{bmatrix} x & y & \mathrm{NCC}_{max} \end{bmatrix}$
end
for *each three adjacent slices* **do**
  compute average NCC_{max}
end
shortlist adjacent slices with highest average NCC_{max}
calculate y_{best} as the median of the y-positions of the shortlisted slice vectors
get x_{best} from the vector $\mathbf{v}(z)$ containing y_{best}
set (x_{best},y_{best},z_{best}) as center of VOI
crop out VOI with a size of $130 \times 130 \times \mathrm{d}_z$ mm^3

4.4 Normalization

The fourth pre-processing module is the image normalization. It is a process to transform all image pixel intensities to a similar range of values. In the deep learning field it is a technique to standardize the input data and it can help models to converge faster. There a several popular normalization techniques, including *feature scaling*, *mean normalization*, and *standard score*. The latter, known in statistics as the *z-score*, is the most widely used among various machine learning (ML) algorithms, such as support-vector machines (SVMs), RF, and ANNs. It's advantage in terms of training a model is that it's mean centering prevents that parameters get abnormally large or become zero during training. Furthermore, it is less sensitive to very high intensities in the image.

The standard score proved to be suitable for the data, the techniques, and the tasks of this work. The mathematical definition is as follows:

$$x' = \frac{x - \mu}{\sigma}, \tag{4.2}$$

where x' is the standardized and x the original pixel value. The standard score normalizes the image pixel values to zero mean ($\mu = 0$) and one standard deviation ($\sigma = 1$). Generally, all images are normalized using the mean and standard deviation of the training data. This proved to be a problem for the knee MRIs since they were acquired from different MR-scanners and exhibited large differences in pixel intensities. To overcome this issue, each 3D MRI was normalized on it's own mean and standard deviation.

4.5 Gold-Standard Segmentation

Gold-standard segmentations are required to quantitatively evaluate any segmentation method (chapter 5). The process of acquiring them is denominated as *manual segmentation* since a substantial part is performed by hand. The user has to manually assign a specific label to the structures of interest in the image. Typical values of such labels for binary segmentation are 1 or 255 for the structure of interest and 0 for the background. Frequently, established edge- and region-based methods [144] are used to generate an initial segmentation and subsequently reduce the amount of manual work. However, the generation of gold-standard segmentations is a time-consuming process.

The manual segmentation of each 3D knee MRIs of this work required three hours on average. Therefore, only a subset of the complete dataset of this work was manually segmented. It consists of 100 coronal MRIs from *Dataset A* and *Dataset B*, which were the first datasets available during the execution of this work. Additionally, 25 sagittal MRIs from all three datasets were manually segmented. Less sagittal than coronal gold-standard segmentations were required, since learning based on coronal images was applied to sagittal ones (sub-section 5.7). The manual results were approved by an experienced radiologist.

A segmentation tool with a graphical user interface (GUI) was implemented to perform the manual segmentation. It was written in C++ using libraries from ITK

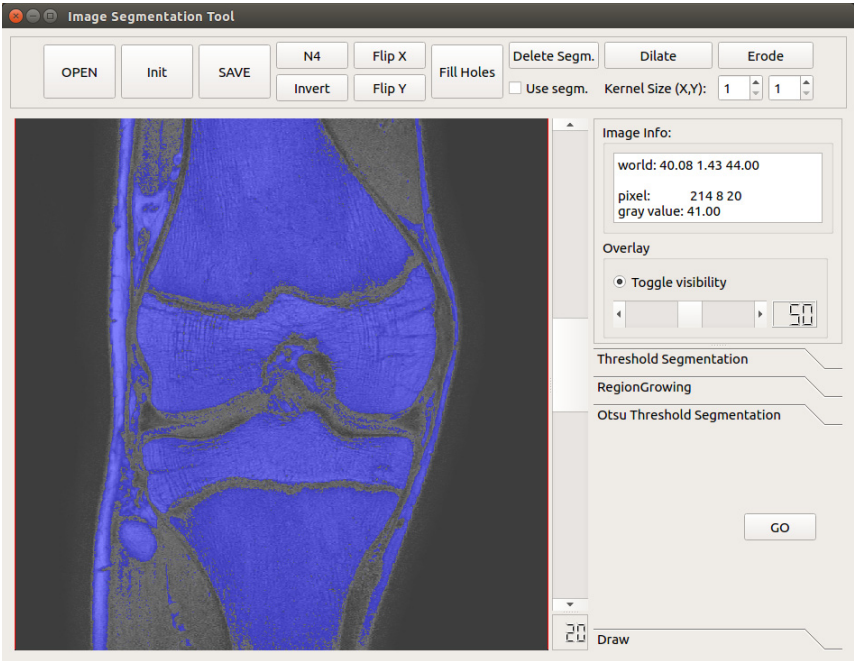

Figure 4.9: The image segmentation tool was designed for the semi-automatic generation of *gold-standard segmentations* of knee bones. It included basic segmentation algorithms, here *Otsu Threshold*, with the option to perform additional automated and manual corrections.

and VTK and the GUI was designed using Qt[11]. The tool allows the user to choose between three established segmentation methods: *Threshold*, *Region Growing*, and *Otsu*. The results can be used as an initial segmentation (Fig. 4.9). Subsequently, the initial segmentation can be improved using dilation and erosion, hole filling, inverting, and flipping methods. To fine-tune the segmentation, the user can perform manual corrections using a draw option. Finally, the resulting binary segmentations can be saved to a hard drive.

In a post-processing step the binary segmentations were transformed to a *label map*[12] which has a different label for Femur, Tibia, and Fibula. First, the connected components in the binary image were detected and then a collection of label objects was generated. Lastly, the labels of the objects were modified such that Femur,

[11] `https://www.qt.io/`
[12] `https://itk.org/Doxygen49/html/classitk_1_1BinaryImageToLabelMapFilter.html`

Tibia, and Fibula had the labels 1, 2, and 3, respectively. In case of performing segmentation separately per bone, another method was implemented to combine them to a label map by requesting the three corresponding labels.

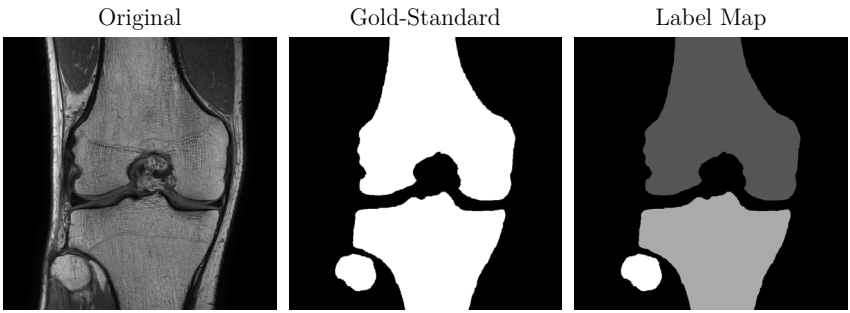

Figure 4.10: Gold-standard segmentations (middle) of the 3D knee MRIs (left) were acquired semi-automatically to evaluate the segmentation approach in the next chapter. Label maps (right) were generated to differentiate the three knee bones.

5 CNN-based Segmentation

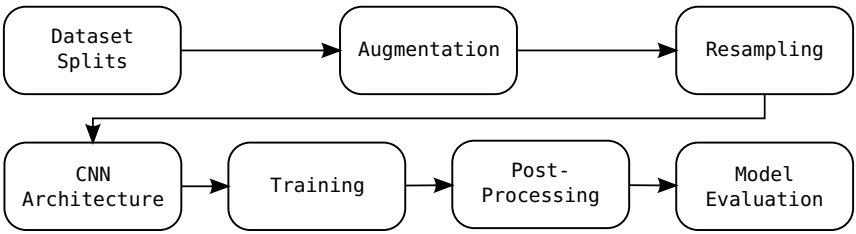

Figure 5.1: CNN-based segmentation used for the bone detection in knee MRIs

Due to the high complexity of the individual structure of growth plates, especially at the edges, and the occurrence of structural-induced intensity variations, established edge- and region-based approaches [144] did not provide adequate detection results. More advanced atlas- and registration-based [76, 85] or fuzzy clustering [138] techniques were only satisfactory to a certain extent. These methods even showed limitations when applied to larger and "less complex" structures such as bones. Especially for image slices containing sparse bone information, the noise rate was high and the detection rate low.

Recently, novel methods in AI, particularly the ones based on deep learning, have been achieving very good results in the segmentation and classification domain. CNNs, first introduced by Yann LeCun [113] in 1990, are nowadays considered state-of-the-art in image semantic segmentation and have been successfully applied on medical images [157].

The established "U-Net" by Ronneberger et al. [157] was the starting point for the development of a CNN-based segmentation of the bones in 3D knee MRIs. U-Net is part of the *autoencoder* models which at first spatially compress input images and then expand them back to their original size. Using this concept, image features can be captured at different spatial resolutions which is particularly useful to detect complex structures.

The workflow of the CNN-based segmentation is depicted in Fig. 5.1. The first step is the *dataset split* (section 5.1) to divide the image data into subsets used for

the training and evaluation of the CNN. Next, the number of training samples is increased with an approach known as *augmentation* (section 5.2) and subsequently all images are brought to a common size with *resampling* (section 5.3). Afterwards, an optimal *CNN architecture* (section 5.4) for the segmentation task is elaborated through multiple test runs. Following the design of the architecture is the actual *training* (section 5.5) of the CNN and *post-processing* (section 5.6) of the output. The final step of the workflow is the *model evaluation* (section 5.7) to compare the predicted segmentation maps by the trained models with the gold-standard segmentations (section 4.5). The entire workflow is performed on knee MRIs in coronal orientation and then adapted and applied to the sagittal data.

5.1 Dataset Split

The inputs of the CNN are knee MRIs and their corresponding gold-standard segmentations (section 4.5). The data is split into three subsets which is a common procedure when using CNNs. The sets are denominated *training set*, *validation set*, and *test set*, and each one is used for a different purpose.

The *training set* is commonly the largest of the three sets and contains the data that is applied to the actual learning process of the neural network. It is the only portion of the data the network will see and directly learn from. The *validation set* is used to regularly measure the performance of the model. If the accuracy on the training set is above the one on the validation set during several consecutive training iterations, than the model is beginning to memorize the training data itself. This phenomenon is known as *overfitting*. Additionally, the validation set is used to fine-tune the model's hyperparameters. By doing so, the network learns from the validation set, but only indirectly. The third and last subset is the *test set*. Contrary to the other sets, it is never involved in the learning process but used as a final performance measure of the model. The objective is to acquire a reliable and independent evaluation of the model based on this third subset.

One-hundred coronal MRIs were split into 70% for the training set, 15% for the validation set, and 15% for the test set. This is a common split ratio of machine learning and deep learning algorithms. The split was random and it was performed for separately per dataset to ensure that all sets had images from *Dataset A* and *Dataset B* (Table 5.1).

Table 5.1: The coronal MRIs were split into three sets for the segmentation task: *training*, *validation*, and *test* sets.

Set	Split ratio	3D images		2D image slices	
		Dataset A	*Dataset B*	*Dataset A*	*Dataset B*
Training	70 %	12	58	492	1392
Validation	15 %	3	12	123	288
Test	15 %	3	12	123	288
Total	100 %	18	82	738	1968

5.2 Augmentation

Data augmentation is a common approach in ML to artificially increase the number of datasets [106]. Generally, the dataset sizes for CNNs range from tens of thousands to millions [106], which is a number difficult to attain for medical applications [157]. The larger the dataset the better the model can account for variability in the data. The types of augmentation on images include translation, rotation, reflections, blurring, brightness, contrast, and gamma adjustments, and other modifications [117, 189]. Augmentation allows the model to learn invariance to such modifications [41, 157]. Moreover, it serves as a type of regularization mechanism to mitigate overfitting by introducing more variation in the data [106, 117, 158]. This approach is generally only performed on the training set from which the model learns directly, while the validation and test sets are excluded in order to properly evaluate the model on real life scenarios.

The following types of augmentation were applied to each 2D image slice of the training set (Fig. 5.2):

- *Horizontal translation* in pixel $\in \{-24,24\}$
- *Rotation* around z-axis in degrees $\in \{-5,5\}$
- *Horizontal flip*
- *FOV change* with a factor $\in \{0.9,1.0,1.1\}$

The augmentation was performed on the downsampled (448×448 pixels) and pre-processed images (chapter 4). Each modification was applied *one at a time*, e.g. an image was not rotated and additionally translated, with exception of the FOV change. The latter was performed on the original images as well as on the translated, rotated, and flipped ones. In contrast to the regular use of augmentation, the FOV change

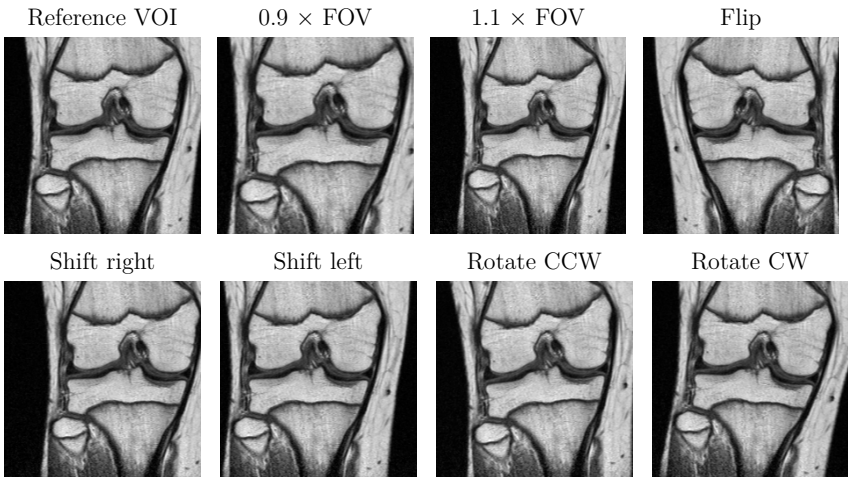

Figure 5.2: Augmentation of knee MRIs to virtually increase the size of the training set used for CNN-based segmentation

alone was also applied to the validation and test sets because it did not change the positions of the knee bones relative to the VOI. By augmenting the images the CNN learns to be invariant to these modifications.

Horizontal translation was introduced to mimic small localization errors of the automated cropping method (section 4.3). Thus, the CNN learns to detect bone irrespective of it's exact positioning in the image frame. The translations were limited to 24 pixels in each direction to maintain the integrity of the shape of the bones. Vertical translation was not performed in order not to undo the effect of automated cropping, which had the advantage of vertically aligning images around the knee joint cavity.

Clockwise (CW) and counter-clockwise (CCW) *rotations* of 5° around the z-axis were applied to the images to account for variations of subject positioning in the MR-scanner and for anatomical differences. Higher rotation amounts were avoided to create rather "realistic" representations of knees in MRI.

Horizontal flips were included to artificially generate images from both the left and right knee. This type of augmentation was especially useful for the robust detection of the Fibula since the bone lay in the bottom left *or* right of the image slice

depending on the laterality. Vertical flips do not represent valid representations of a MRI sequence of the knee and were consequently not considered.

FOV change were brought into use to account for anatomical differences in subjects. The center of the standardized VOI (section 4.3) was used to create two additional VOIs; one 10% smaller and the other 10% larger than the reference size of 130×130 mm^2. This type of modification effectively created three different FOVs of the same MRI (Fig. 5.2).

Applying translation and rotation after the automated cropping (section 4.3) causes parts of the structures to exit and new pixels to enter the image frame. These new pixels are unknown information in the VOI and filled with zeros, but do not represent actual anatomical structures. Yet, this information is, in most cases, available in the full size images. Therefore, augmentation was performed *prior* to the actual cropping to recover "lost" anatomical structures. This was an improvement in comparison to the initial approach in [148]. Refer to appendix C for more details and visual examples.

5.3 Resampling

In theory, CNNs have no image size limitation, but it is recommended to reduce the spatial resolution to decrease the number of calculations. Moreover, all images are required to have equal dimensions for the training. Thus, the preprocessed, augmented, and cropped images of this work were resampled to a uniform size per slice of 224×224 pixels, while also being a common resolution used for CNNs.

Resampling can have the drawback of causing loss of information, i.e. detail loss of anatomical structures. Nevertheless, only a small amount of resampling was necessary due to the automated cropping (section 4.3), which previously reduced the image size. The final size of 224×224 pixels per image slice provided enough detail to identify the shape of Femur, Tibia, and Fibula. The images of all three sets were resampling to this size since it is a requirement for the input shape of the CNN.

5.4 CNN Architecture

A CNN is composed of an input and an output layer and multiple hidden layers in between, similarly to the concept of a *multilayer perceptron* (MLP). The data is passed from the input layer to the hidden layers, where the non-linear transformations of the data are computed, and finally one or more outputs are predicted in the output layer (Fig. 5.3). Typically, the hidden layers of a CNN include convolutional layers, activation functions, pooling layers, regularization layers, and fully connected layers. Choosing the correct amount, the order, and the combination of layers requires experimentation and is dependent on the task and the input data.

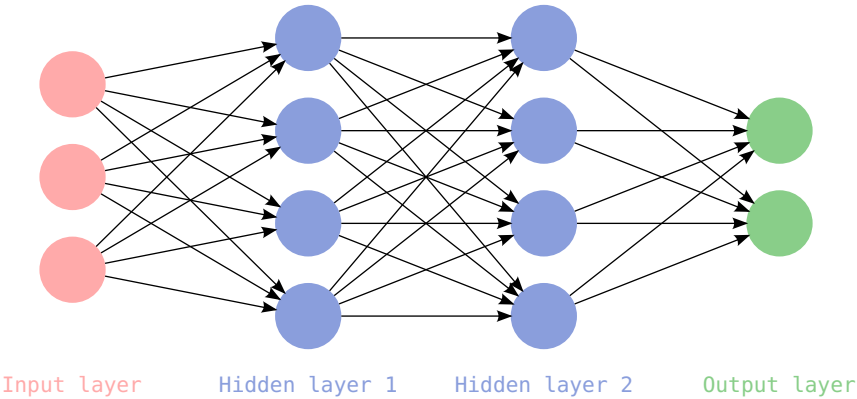

Figure 5.3: A multilayer perceptron

The starting point for establishing the architecture for this work was U-Net [157] which has been successfully applied to segment biomedical images and has been setting state of the art results (Fig. 5.4). Encoder-decoder models such as U-Net are the de facto standard in the field of image segmentation [4, 19, 20, 25, 120, 132, 139, 157, 229]. These models processes the data along two paths: the *contracting path* (encoder) spatially compresses the images while the *expanding path* (decoder) upsamples the images to restore the original size. This allows the network to capture features at different spatial resolutions while also increasing the processing speed due to downsampling.

Next, an overview of the layers relevant for this work, followed by the designed architecture for the segmentation task of this work.

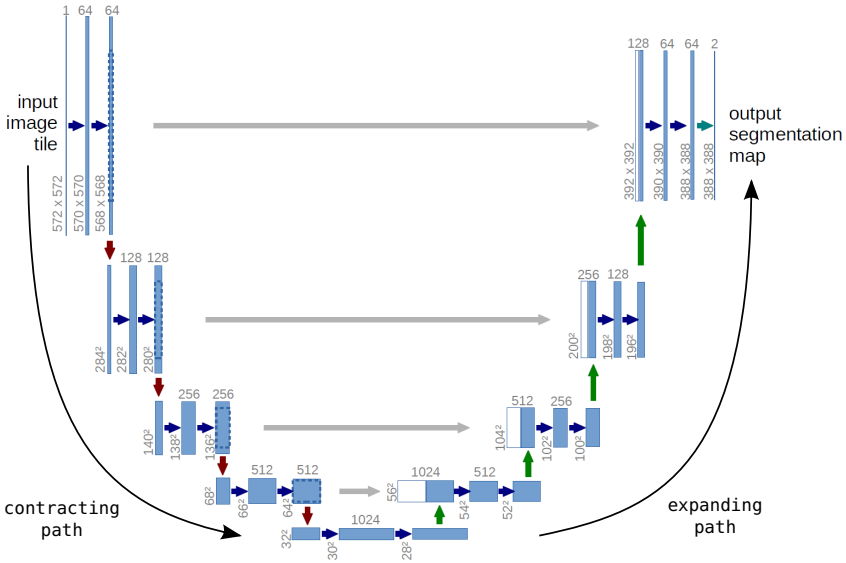

Figure 5.4: U-Net, a popular CNN architecture for segmentation, first compresses the input data and then expands it back to it's original size (adapted from [157])

Convolutional layer. The primary building block to construct the architecture of a CNN is the convolutional layer [73] and therefore the term *convolutional* in "CNN". Convolution is about multiplying the value of each image pixel and it's neighbouring ones with a *kernel* and then summing them up to calculate a new value at the pixel's location (Fig. 5.5). For CNNs, the kernel is generally a matrix with size 3×3, 5×5, or 7×7. The kernel is applied across the entire image in a sliding window fashion to generate a feature map. Each convolutional layer is composed of multiple kernels and the number typically becomes larger in deeper layers. The goal of CNNs is to learn the values of all kernels, i.e. the *weights*, to produce the best results for the given task. Furthermore, by using two or more convolutional layers in succession, a CNN is able to learn complex and hierarchical features [23, 73]. Recently, alterations of the regular convolution operation have emerged such as *depthwise separable* and *dilated convolutions*. Depthwise separable convolutions have been used to create smaller and faster models with comparable performance [23, 24], especially useful for mobile devices and real-time applications [227]. Dilated convolutions on the other hand, have gained popularity since they enable to capture features at multiple scales without the need of scaling operations [134, 146, 221, 223].

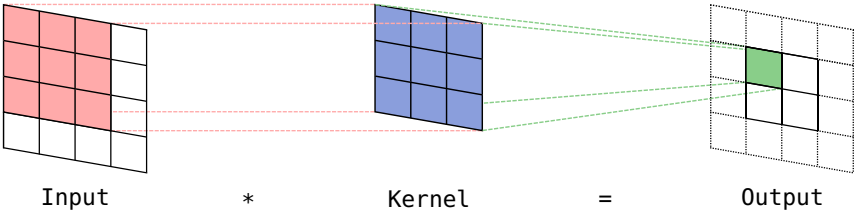

Figure 5.5: Convolution is about sliding a kernel across the image to generate a feature map to be used as input for other CNN layer

Activation function. Activations functions are placed between the layers of a CNN, more precisely after convolutional layers, to introduce a non-linearity. The purpose is to generate features that are non-linear transformations of the input [73]. Otherwise, a simple linear transformation could be calculated and the benefit of building a deep network would vanish. The most popular activation function in modern neural networks is the *rectified linear unit* (ReLU) [137]. Other variants of this activation function are PReLU [74], LReLU [124], and ELU [26]. They all return the identity for positive arguments but handle negative ones differently (Fig. 5.6). The advantage of ELU is that its exponential part for negative arguments pushes the mean activation to zero, which accelerates the learning process and reduces the bias shift passed from the activation function to the next layer [26].

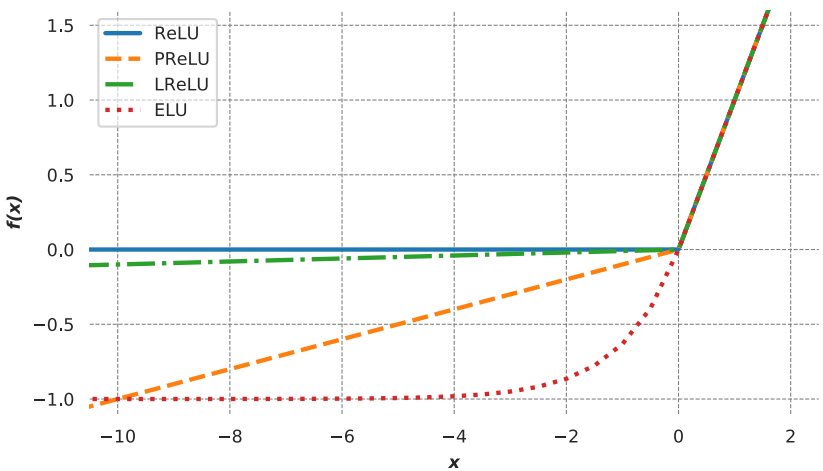

Figure 5.6: Common activation functions of neural networks

Pooling layer. Pooling is an operation to reduce the size of the output feature maps of convolutional layers. The benefit of this operation is that it introduces spatial invariance [73, 147, 166], reduces the number of calculations, and reduces the possibility of overfitting [147]. One can choose between *max pooling* and *average pooling* of which the first one is more common and generally performs better [166]. Max pooling computes the maximum value in each subregion of the feature maps and sets the result in the output (Fig. 5.7). If the max pooling filter size is set to 2×2 and the *stride*[1] parameter is set to 2 then one can effectively reduce the feature values by half. By combining convolutional and max pooling layers one can reduce the input to a dense representation of the most important features.

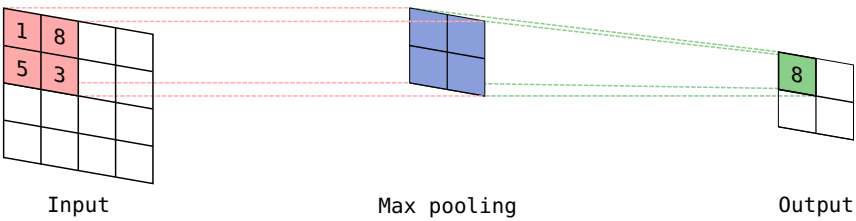

Figure 5.7: Max pooling (2×2; *stride* $= 2$) computes the maximum of a 2×2 subregion of the input and halves the input size

Skip connections. Each max pooling layer reduces the size of the image and causes loss of spatial information. To overcome this issue, *long skip connections* can be introduced to pass spatial information from the contracting to the expanding path of the network [44, 88]. Another type of skip connections are the *short skip connections*, also known as *residual connections*, which can be added around one or more layers. They are used to address the issue of *vanishing gradient*[2] when building very deep networks [44]. Moreover, their integration in segmentation models has shown that the model converges faster during training, it's optimization is easier, and it's segmentation accuracy improves [75, 88].

Regularization. Many different approaches can be used to address the issue of overfitting mentioned in section 5.1. A popular and commonly used technique is

[1] Step of the convolution operation in pixels

[2] The kernel weights of the convolutional layers are modified during training by the partial derivative of the error function. Especially in very deep models, the derivative can become vanishingly small, preventing the weights to be effectively changed [69, 81, 82].

Dropout [194] (DO) which randomly drops kernels and their connections to other layers during training. As a consequence, the model is encouraged to learn independent features since one kernel cannot rely on the effect of other ones [73, 194]. Other types of regularization are *L2* and *L1* regularization, which are applied to the kernel weights to reduce overfitting [73]. *Batch Normalization* [87] (BN) is yet another technique which normalizes the inputs of a layer and can induce faster training and higher robustness of network weights initialization [26, 83, 87].

Final architecture and hyperparameters

The final architecture for the segmentation task of this work is depicted in Fig. 5.8. The input is a 2D MR image of size 224×224 and is first downsampled along the contracting path to the left, then upsampled on the expanding path to the right, and finally the output segmentation map is generated. The spatial information lost during downsampling is recovered on the expanding side using skip connections. The final architecture resulted from a rigorous analysis of multiple variations of the network hyperparameters described in the previous paragraphs, e.g. different number of kernels per convolutional layer, increments of kernels with model depth, activation functions, regularization techniques, and so on.

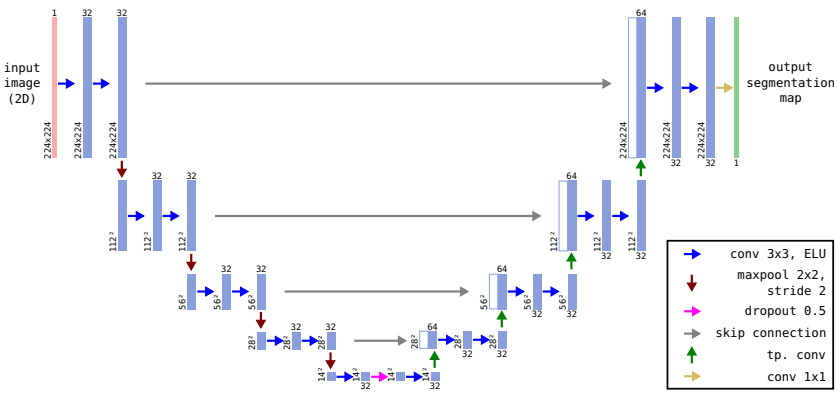

Figure 5.8: The final architecture for CNN-based segmentation of knee MRIs

The main building blocks of the architecture are the *Down-Block* for the contracting path, the *Up-Block* for the expanding path, and the *Bottom-Block* which connects both paths (Fig. 5.9):

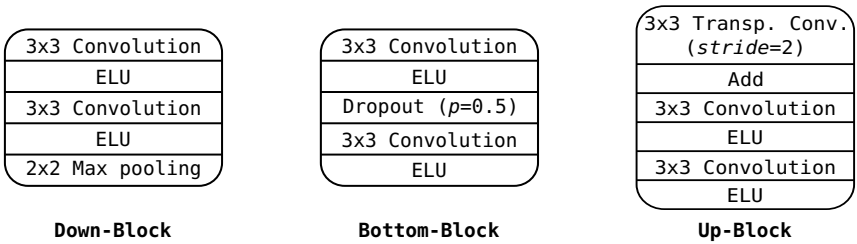

Figure 5.9: The building blocks of the CNN for segmentation

The *Down-Block* features the following sequence of units: Convolution → ELU → Convolution → ELU → Max pooling. Two convolutional layers are used in succession to learn complex and hierarchical features. Each convolutional layer is followed by an ELU activation function to introduce non-linearity. It gave the best results in comparison to ReLU, PReLU, and LReLU. The convolutional layers of the CNN are followed by a max pooling layer with a filter size of 2×2 and a stride of 2, which reduces the size of the feature maps by half. It does not only reduce computational power but also create a dense representation of the most important features. The depth of the network is set to 4, i.e. a total of four Down-Blocks on the contracting path and four Up-Blocks on the expanding path. With each additional depth the network captures features at a different spatial resolutions. Reducing the depth below 4 resulted in lower segmentation scores and setting the maximum depth of 5 (downsampling limit of the image) did not improve the performance either. All convolutional layers in the network have 32 outputs. Training a higher number of features per layer or incrementing the number of outputs with network depth did not improve the segmentation quality and rather contributed to overfitting. The convolutional kernels have a size of 3×3. Higher kernel sizes of 5×5 or 7×7 resulted in lower performance.

The *Bottom-Block* connects the contracting and expanding paths. It is composed of the following units: Convolution → ELU → Dropout → Convolution → ELU. It operates on the lowest resolution of the input image, i.e. 14×14 pixels. Here, the convolutional layers have the same settings as in the Down-Blocks. The key unit is

Dropout. This technique encourages the model to learn representative and independent features. This type of regularization was successful in preventing overfitting.

The *Up-Block* operates on the expanding path of the network and features the following sequence of units: Transposed convolution → Convolution → ELU → Convolution → ELU. Here, the transposed convolution or *deconvolution* reverses the effect of max pooling, i.e. it upsamples the feature maps by a factor of 2. The spatial information from the contracting path is transferred via *skip connections* and appended to the outputs of the transposed convolution layer resulting in 64 feature maps. Afterwards, a sequence of two convolutions with non-linearities generates further feature maps. The convolutional layers and the activation functions have the same properties as the ones of the Down-Blocks. A total of four Up-Blocks are set up to restore the full extent of the input image of 224×224 pixels.

The last part of the network is the generation of the output segmentation maps. This is done by applying a final convolutional layer with a kernel size of 1 to the output of the last Up-Block. This effectively reduces 32 feature maps to 1 segmentation map. The activation function for the last convolutional layer is the sigmoid function which outputs a probability between 0 and 1 for each pixel. The higher the value, the more likely the pixel is a bone structure. The final binary segmentation is acquired by setting a threshold probability. The optimal resulted in 0.3.

The objective of the CNN is to learn the values of each convolutional kernel, i.e. the *weights* of the network, along with the *biases* to generate feature maps:

$$Y_j = \sum_{k=1}^{32} X_{jk} \cdot W_k + B_j, \tag{5.1}$$

where Y_j is the j-th output of the convolutional layer, i.e. the feature map, $\{k \in \mathbb{N} : 1 \leq k \leq 32\}$ the kernel number, X_{jk} the j-th input of the layer, W_k the weights, and B_j the j-th bias. The final network architecture for the segmentation task has a total of 194.561 trainable parameters given the convolutional and deconvolutional layers. It is approximately 150 times smaller than U-Net which has 31m parameters.

5.5 Training

The training of a neural networks resembles the way humans learn. When we perform a task which is unknown or very difficult, we merely guess the action to get the best result. Comparing our guess with the actual result is how we learn and improve in the future given similar scenarios.

Similarly, a neural network learns from the given data during the training in two phases: *forward* and *backward propagation* (Fig. 5.10). In the first phase, the input data is fed forward through the network and a prediction is made. At first, like humans, it is a random guess since ANNs are initialized with random values. In the second phase, the prediction of the model is compared to the ground truth. Finally, the error is calculated and fed backwards through the network to optimize the weights of the model.

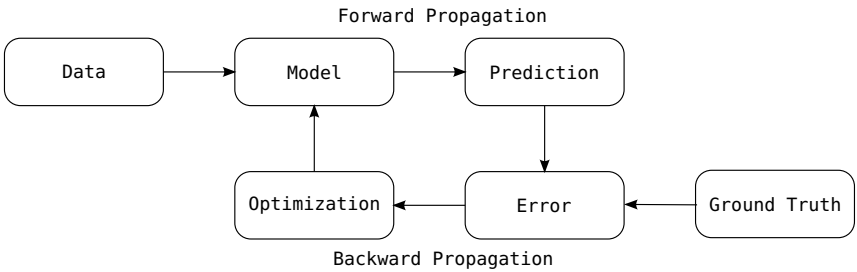

Figure 5.10: The iterative process of forward and backward propagation during the training of a neural network

The training data is fed into the network in batches. Each batch contains a number of random samples, which is referred to as the *batch size*. One full iteration over all training samples is defined as an *epoch*. Generally, models are trained over multiple epochs to reduce the error between the prediction and the answer. By providing the data in random batches the network encounters a new scenario in each epoch, which can help it generalize better. Typical batch sizes are in the range of 32 to 512 [97]. Earlier works, e.g. [97, 114], suggested that large batch sizes lead to a generalization gap. A more recent work [83] says that the generalization ability of models using large batch sizes can be good-as or even better than the one with a small batch size, provided the necessary adjustments to the training hyperparameters. The optimal batch size for the segmentation network of this work was 48. The number of epochs

was set to 50. The *early stopping* technique was used to terminate the training process as soon as the model stopped improving on the validation data. This is done to prevent overfitting of the model on the training data. The *patience* is the parameter of early stopping to be set and represents the number of epochs to wait for the model to improve on the validation data. The optimal patience value for the segmentation task was 9.

To compute the error between the prediction and the correct answer, a metric and loss function have to be defined. An optimal metric for the segmentation task is the *Dice Similarity Coefficient* (DSC) which is also known as $F_1\,score$:

$$F_1\,score = \frac{2TP}{2TP + FP + FN} \tag{5.2}$$

where TP are true positives, FP false positives, and FN false negatives.

Another metric that can be taken into consideration is the cross entropy, which is popular for classification problems. It is applicable to segmentation tasks as well since a segmentation can be seen as a classification for every output image pixel. Ultimately, the $F_1\,score$ was selected as the metric since it delivered better results. The $F_1\,loss$ function is derived from the selected metric and accounts for the continuous probabilities resulting from the sigmoid activation function of the output layer:

$$F_1\,loss = 1 - F_1\,score \tag{5.3}$$

To optimize the parameters of a deep learning model, the loss is commonly minimised or maximized through a process called *stochastic gradient descent* (SGD) or ascent. It takes a random batch of the training set and iteratively adjusts the weights in small steps. Over the years, different variants of SGD were introduced to the field. The *Adam optimiser* [98] is such a variant, which enhances SGD amongst other things by using what is called an "adaptive momentum estimation". It adaptively adjusts the *learning rate* which defines how much the weights are changed in one training step. Adam is a robust optimizer that has been successfully used in segmentation tasks of medical data [21, 40, 96, 163, 192, 202]. Thus, it is selected for the optimization process in this work. The main parameter to be defined for Adam is the *learning rate policy*, which represents the amount the learning rate is change from one epoch to the next. Test runs indicated that an initial learning rate of 0.001 without decay was optimal for the given network setup and task.

5.6 Post-Processing

The output segmentation maps occasionally exhibited small and falsely detected areas outside the bone structure and small holes inside the bone structure. Therefore, post-processing on the predicted segmentations of the CNN was necessary to correct minor false predictions. While numerically the post-processing does not make a significant impact on the segmentation performance, qualitatively it improves the segmentation maps (Fig. 5.11).

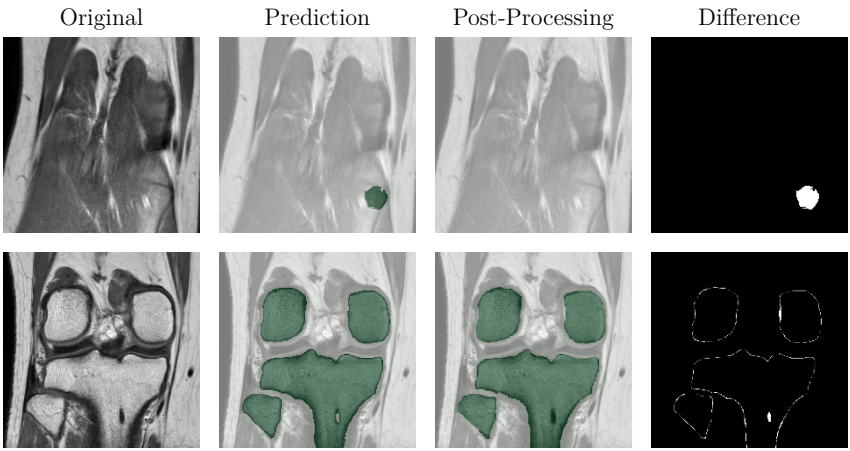

Figure 5.11: Post-processing to enhance the results of the CNN-based segmentation. The first row shows a wrap-around artefact which is partially mistaken as bone. In the second case a dark blob-like spot inside the Tibia is not detected. The post-processing can correct such minor prediction errors.

Morphological operations (erosion and dilation) are used to remove small structures which were not connected or related to any of the three knee bones. A method named "binary opening" [116] removes small objects by applying an erosion first, followed by a dilation on the binary image. An implementation[3] exists in the Python library "SimpleITK" and could successfully be integrated into the workflow of this work. The fundamental parameter of the method is the *kernel radius*, which defines the size of the objects to be removed. The radius was only adjusted in the z-direction

[3] `https://itk.org/SimpleITKDoxygen/html/classitk_1_1simple_1_1BinaryOpeningByReconstructionImageFilter.html`

to 4, meaning that objects were removed if not connected over at least 4 adjacent slices of a 3D knee MRI. Higher radii were not feasible since it risked the removal of the Fibula, which only appeared in a few image slices.

Small holes inside bone structures are filled using "binary closing" [116]. Opposite to the opening method, it first dilates and then erodes connected components in the image. An implementation[4] in Python exists as well and was included as a further post-processing steps of the knee segmentations. The binary closing is performed on each 2D image slice providing a kernel radius of 4. Thus, only holes with a diameter smaller than nine pixels are closed.

5.7 Model Evaluation

Segmentation Models

The gold-standard segmentations includes a different label for each bone. This enables the possibility to train three types of segmentation models:

- *Merged model*
- *Single bone model*
- *Combined model*

For the *merged model* the labels of each bone are merged to a single channel, i.e. all bones pixels have the value 1. For the *single bone model* each bone is used separately to train a model. For the *combined model* the predictions from each separate bone model are combined to one segmentation.

Model Exploration

Neural networks are frequently considered black boxes since it is difficult to extract and present the learned features and the decision made in a human-readable manner [23]. Fortunately, two factors can help to get a better understanding of a segmentation network. Firstly, the segmentation task as such is an image to image pipeline. This makes it easier to deduce how the input was mapped to the output since both

[4] `https://itk.org/SimpleITKDoxygen/html/classitk_1_1simple_1_1BinaryClosingByReconstructionImageFilter.html`

are in the image domain. Secondly, CNNs mitigate most of the black box reasoning because the learned features can be visualized and are based on the concept of the visual cortex of the brain [23].

To explore the model and get a better understanding of the learned features for the segmentation task, the individual kernels and feature maps are extracted and visualized. The weights of the desired model layer can be extracted using a function[5] of "Keras"[6], a high-level neural networks API utilized in this work. Intermediate feature maps of the model are generated with the following two steps: first, a reduced version of the trained model is created by setting it's output to the output of the desired layer, and second, an image slice is fed to the reduced model to acquire the feature maps at the desired layer.

Noise Exploration

The gold-standard segmentation can be subjected to noise since unmeant errors can occur during their generation. Moreover, it may be unclear whether certain areas in the image belong to the target structure or not. Lastly, small errors can be introduced due to the downsampling of the data as part of the image pre-processing.

To evaluate the network's robustness against noise, another training environment is set up: synthetic noise, in form of erosion and dilation with a kernel size of 7, is applied to the gold-standard segmentations of the training and validation sets. The evaluation is then performed on the *unmodified* test set.

Transfer Application

Neural networks are generally trained on a large amount of data and perform well when presented with data that lies in the range of variation of the training set. Nevertheless, they become unreliable when applied to related but different data. For this purpose, CNNs prove to be more useful. Due to the sliding window fashion of convolutions they learn local patterns in the image [23]. Thus, CNNs have the advantage that they learn features that are translation invariant and can therefore recognize similar patterns anywhere in the image [23]. To test this attribute of

[5] `https://keras.io/layers/about-keras-layers/`
[6] `https://keras.io/`

CNNs, the merged model, trained on cropped MRI slices in coronal orientation, is applied to unknown data in two different scenarios.

The first test scenario includes MRI slices which are preprocessed with BFC and normalization but are *not* cropped. Therefore, they have a much larger FOV than the data in the training set and their size is 448×448 instead of 224×224. The implemented network architecture (section 5.4) does not include any dense layers and is consequently able to process image data of any size.

The second test scenario incorporates MRI slices in sagittal orientation. The preprocessing steps match the ones in the first scenario. In this case the model is confronted with data with much larger differences to the training set since the knee bones are represented in a different orientation.

In the third and last scenario the concept of *transfer learning* is used. It is about applying the learned knowledge gained on one problem to solve a similar but different one [151]. For this purpose, the learned weights from the merged model were used to initialize a new model trained on a smaller number of gold-standard segmentations of sagittal MRIs, instead of applying the merged model on sagittal images directly.

Transfer Learning

The setup for the sagittal model did not vary significantly compared to the merged model used for coronal MRIs. For this third scenario, a total of 25 sagittal MRIs were manually segmented, including samples from all three datasets (section 4.5). This approach requires less ground truth data since transfer learning is used. The preprocessing was similar to the one on coronal images except that more augmentation in form of translation and rotation was applied to the training set. Finally, the training set included approximately 24k sagittal 2D images in comparison to the 33k generated for the coronal merged model. The network architecture was not modified at all. The key part was the training process: the weights of all network layers were initialized with the values from the learned weights of the coronal model. The optimizer, the learning rate, the loss function, and the batch size were identical. All layers were then retrained on the sagittal data for only five epochs without early stopping. Alternatively, only a few network layers can be retrained with a reduced learning rate in order to fine-tune the model for the new data. This approach did not prove to be better in comparison to pure initialization of learned weights. Lastly,

the model was evaluated using *leave-one-out cross-validation* (LOOCV), i.e. each of the 25 sagittal images represented the test set once. The post-processing step was not modified in comparison to the coronal case.

2D vs. 3D CNN

CNNs are commonly used on 2D data such as photos but can be applied on data of any dimension. For exampled they can be used to analyze 1D data such as sentences for Natural Language Processing (NLP) or signal data for audio processing. In the medical field, data is often volumetric. To exploit the 3D context, CNNs can include 2D convolutions for each image slice or 3D convolutions for the whole image volume. Both approaches have been used in many applications on medical data [18, 25, 37, 40, 42, 43, 95, 99, 132, 147, 148, 159]. While a few studies on image segmentation achieve better results using 3D CNNs [25, 37], other studies [147, 148] report higher performance using 2D CNNs.

The volumetric knee MRIs of this work enable the possibility to design a CNN that can handle inputs is 3D. Moreover, it allows the comparison of this new architecture with the proposed 2D segmentation network (section 5.4) in order to provide the community with the results based on knee MRIs.

The only change that is necessary in the pre-processing (chapter 4) is to set all volumetric knee data to the same number of slices. For the 3D images of *Dataset A*, empty slices are added to the front and back to obtain 48 slices and then every second slice is selected to match the 24 slices of the images from *Dataset B*. The downside of the 3D approach is that data is "thrown away". Finally, the input data of the 3D CNN is a volume of size $24 \times 224 \times 224$.

A few experiments are then performed to define the optimal architecture of the 3D CNN for segmentation, including the variation of the number of kernels, the increment of kernels with network depth, the inclusion of Batch Normalization and Dropout between convolution layers, and batch sizes. The resulting 3D network is very similar to the 2D one. The depth, the number of kernels, and their increment with depth is unchanged. The main modifications includes 3D instead of 2D convolutions and up-sampling instead of transposed convolution. The final building blocks for the 3D CNN for this work are depicted in Fig. 5.12.

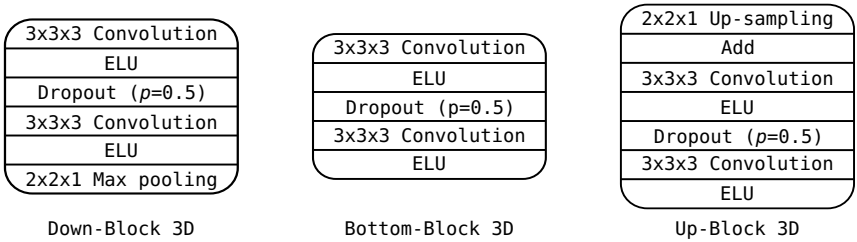

Figure 5.12: The building blocks of the 3D CNN for segmentation

The Down- and Up-Blocks include Dropout between convolutional layers due to increased overfitting observed in the 3D case. Max pooling ($2 \times 2 \times 1$) is not applied to the z-axis of the 3D image data due to the low number of slices of the MRIs in comparison to their in-plane resolution. Similarly, the up-sampling is only applied to the slices and not the z-axis. The training setup is identical as in section 5.5, but the batch size is reduced to 2 due to the increased memory requirements of volumetric data.

Model Performance

Typical measures used for the evaluation of the segmentation performance are the *Dice* and the *Jaccard* similarity coefficients, which are known in the machine learning field as the F_1 *score* and the *Intersection-over-Union* (IoU), respectively. The mathematical formulations are the following:

$$\mathrm{DSC} = \frac{2TP}{2TP + FN + FP} \tag{5.4}$$

$$\mathrm{IoU} = \frac{TP}{TP + FN + FP} \tag{5.5}$$

Two further measures to quantify the quality of the segmentation are the *precision* and *recall*. Precision is the amount of predicted area that is true and recall is the amount of the ground truth that is detected by the model:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.7}$$

Finally, the total error of the model for the segmentation task can be computed with the following equation:

$$\text{Total Error} = \frac{FP + FN}{TP + TN + FP + FN} \tag{5.8}$$

where TN are true negatives.

For a reliable and independent numeric evaluation the trained segmentation models are evaluated on the test sets, i.e. data the model has never seen during training. Furthermore, *5-fold cross-validation*[7] is performed to evaluate the model's robustness and generalizability. In addition, due to the stochastic nature of CNNs, the training is repeated five times for each fold. The final evaluation measures are averaged over 25 training rounds.

[7] *K-fold cross-validation* (CV) is the evaluation of a model on k distinct and independent test sets

6 Age Estimation

This chapter provides the description of three methods for age estimation based on different AI techniques and data (Fig. 6.1). All three methods train both machine learning algortihms on regression (ML-R) of the chronological age (CA) and on classification (ML-C) of the 18-year-limit in a final step.

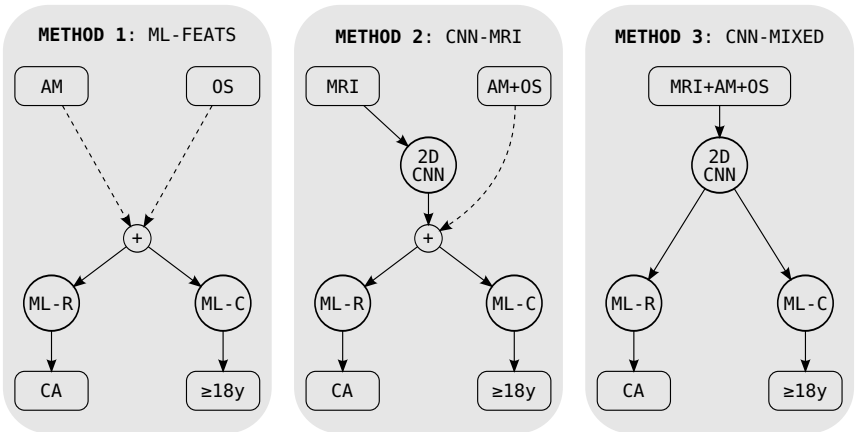

Figure 6.1: Three methods for age estimation of male adolescents and young adults. *Method 1* is purely based on machine learning (ML) using anthropometric measurements (AM) and ossification stages (OS) separately or combined as input features (feats). *Method 2* is based on convolutional neural networks (CNNs) using knee MRIs as input data. *Method 3* is based on CNN as well but handles mixed input data (image and numeric). Ultimately, all methods train both ML algorithms on regression (ML-R) of the chronological age (CA) and on classification (ML-C) of the 18-year-limit in a final step. *Method 2* uses age predictions per knee MRI slice from the CNN as input to ML-R and ML-C and *optionally*, integrates AM and OS as well. *Method 3* handles all data in the CNN and uses age predictions per knee MRI slice from the CNN as input to the ML algorithms.

Method 1 (section 6.1) is purely based on ML algorithms using AM and OS separately or combined as input features (feats). *Method 2* (section 6.2) is based on

CNNs using knee MRIs as input data in a first step. The output of the CNN is a single age prediction for each image slice of the knee MRI volume. Next, the predicted ages of all image slices associated to one subject are combined and used as an input to the ML algorithms. Optionally, AM and OS are incorporated to the regressors and classifiers as additional features to observe their impact on the age estimation performance. Finally, *Method 3* (section 6.3) is based on CNNs for multi-input and mixed data. It combines the numeric data (AM and OS) with the image data (knee MRI) in a single CNN to predict the age per image slice. Then, all age predictions related to one subject are used to train ML algorithms on CA regression and majority classification.

6.1 Method 1: ML-FEATS

Most methods currently used in practice for age estimation infer the chronological age of an individual from the ossification stage of a bone with the use of statistical analyses or using the minimum-age concept (chapter 2). These methods are manual, labour-intensive, subjective, and not suitable to regress the age of an individual.

Hence, motivation of *Method 1* is to reproduce medical methods for age estimation currently used in practice, but based on an automated, reproducible, and learning-based approach. Instead of relying on manual and subjective assessments of age-related characteristics, such as growth plate maturation, by experts, the proposal is to use ML algorithms to learn the relationship between the characteristics and chronological age. ML algorithms are straightforward to use and are fast to train with modern programming libraries.

A further motivation of *Method 1* is that it offers the possibility to corroborate that AM have a large error of margin for age estimation as mentioned in [56, 57, 172]. From a statistical point of view, the weak correlation between AM and the age of the subjects from this work suggests a rather low suitability for the task (Fig. 6.2). On the other hand, a positive correlation between the ossification stage of growth plates and the chronological age of adolescents and young adults has been ascertained in literature (chapters 1 and 2). The OS acquired from the underlying population of this work supports this fact (Fig. 6.3).

Method 2 analyzes multiple ML algorithms on age regression and on majority classification based on AM and OS, separately and combined. It will become apparent,

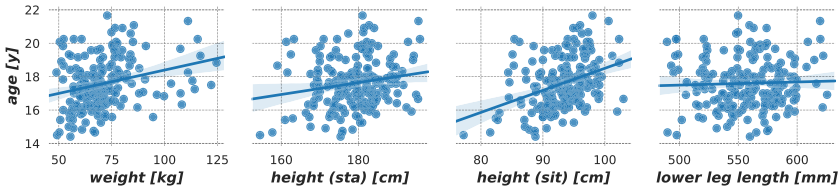

Figure 6.2: Relationship between anthropometric measurements and chronological age of adolescents and young adults. The Pearson correlation coefficients of the measurements were $r = 0.27$ ($p < .05$), $r = 0.19$ ($p = .01$), $r = 0.37$ ($p < .05$), and $r = 0.04$ ($p = .61$) from left to right. The relationship is significant for a significance level of 5% except for the lower leg length.

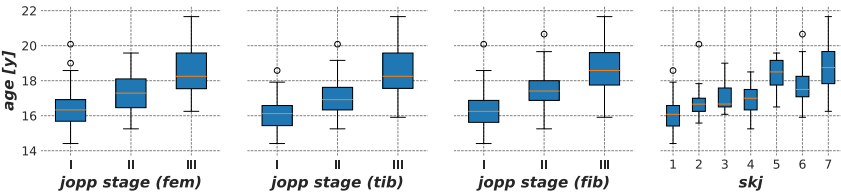

Figure 6.3: Boxplots of chronological age vs. ossification stages in the knee. Three-stage system by Jopp et al. [92]; SKJ is the sum of stages for all three growth plates of the knee. The Spearman correlations of the stages per knee bone and the SKJ with age were $r = 0.68$ ($p \ll .05$), $r = 0.71$ ($p \ll .05$), $r = 0.70$ ($p \ll .05$), and $r = 0.73$ ($p \ll .05$) from left to right.

which data is most suitable for both tasks and whether a combination of the data can reduce the margin of error as suggested in [12, 49, 173].

6.1.1 Data Preparation

The data for this method are AM and OS from *Dataset A* and *Dataset B* (sections 3.2 and 3.4). AM and OS were not available for *Dataset C*. To be comparable to *Method* 2 (section 6.2) and *Method* 3 (section 6.3), only the data from subjects with an MRI examination are considered. This amounts to a total of 185 datapoints which require the following data preparation steps for the ML algorithms: *data cleaning*, *data split*, and *data standarization*.

The *data cleaning* removes the samples from subjects without a knee MRI. Additionally, missing AM are filled using the mean value of the measurement of the respective age group. This process is called *imputation* and is common practice in ML. The last step of data cleaning is to create an additional target variable for the majority classification: all adults ($\geq$ 18 years) are assigned the value 0 and all minors the value 1.

After the cleaning, the data is split into two sets, which is commonly done in machine learning: *training* ($n_{tr} = 150$) and *validation* ($n_{te} = 35$) sets. The training set is larger in size and is the only part of the data that is used to train the ML algorithms. The test set is utilized to evaluate the performance of the trained models on the regression and classification tasks on unknown data. To perform a valid comparison between the three methods for age estimation described in this chapter, the test set includes the same subjects in all cases.

The last preparation step is *data standardization*. The standard score (equation 4.2) is used since many ML algorithms require mean centering and unit variance to effectively learn from all features. This data transformation technique is useful when large disparities between the values of different features are present. For example, the LLL for the subjects of this work were in the range of 500 till 600, the standing height and weight had mostly values under 100, and the OS values below 10. Standarization is performed with an available module[1] from Scikit-learn [145], a popular ML library for the Python programming language. The module computes the mean and standard deviation of each feature separately. Means and standard deviations are always calculated for the training set and then used to transform the features in the test set.

6.1.2 ML Setup

Multiple ML algorithms for age regression and majority classification are used and evaluated using AM and OS both separately and combined. The Scikit-learn library in Python offers a large variety of algorithms and allows the user to tune their parameters for the underlying data and task. The following ML algorithms are selected:

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler`

- Regression
  - ▷ Ordinary least squares Linear Regressor (LR)
  - ▷ Support-Vector Regressor (SVR)
  - ▷ Decision-Trees
    - ◇ Random Forests Regressor (RFR)
    - ◇ Extremely Randomized Trees Regressor (ETR)
    - ◇ Gradient Tree Boosting Regressor (GBR)
- Classification
  - ▷ K-Nearest Neighbours Classifier (KNC)
  - ▷ Support-Vector Classifier (SVC)
  - ▷ Decision-Trees
    - ◇ Decision Tree Classifier (DTC)
    - ◇ Random Forests Classifier (RFC)
    - ◇ Extremely Randomized Trees Classifier (ETC)
    - ◇ Gradient Tree Boosting Classifier (GBC)

Each ML algorithm has multiple parameters that can be adapted to improve the performance on the given task. To find the best parameters of each algorithm for both age regression and majority classification, the following two procedures are undertaken in succession:

1. Gradually increase one parameter while keeping all others constant and observe the change in performance

2. Search the entire parameter space for the best result using cross-validation

The first procedure is performed separately on each important parameter of each algorithm. Individual parameters are changed and the impact on the task-specific metric is evaluated (Fig. 6.4). The importance here is to detect overfitting, i.e. when the performance improves on the training set and deteriorates on the test set given a change in parameter value. An example is given in Fig. 6.4 for two parameters, the *number of estimators* and the *maximum tree depth*, of the GBR[2]. Increasing the number of estimators, i.e. regression trees, of GBR leads to a rapid decline of the error on the training set and slight improvement on the test set. Although the gain on the test set is not as large as on the training set, using more estimators for tree-based ML algorithms generally does not risk overfitting due to

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html`

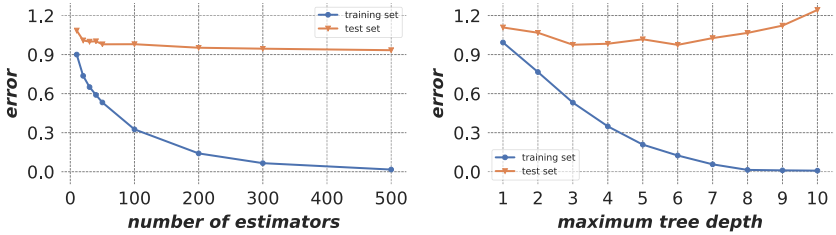

Figure 6.4: Analysis of two parameters of a Gradient Tree Boosting algorithm trained on age regression. An increase of the number of estimators leads to lower error for both the training and test sets (left). Contrary, overfitting is observed when increasing the maximum tree depth beyond 6. The inspection of these plots can aid in the selection of suitable parameter value ranges for further analysis.

their random nature [10]. The maximum tree depth parameter shows a clear sign of overfitting when increasing the value above 6 and therefore lower depths are preferred. Ultimately, the first procedure is not utilized to determine the optimal value of each parameter but instead, to limit the value range of each parameter to be used in the second procedure. Otherwise, there is a higher chance that the selected parameters are only suitable precisely for the given data.

Subsequently, the second procedure is used to perform an exhaustive search over a parameter space with the reduced value ranges. The Scikit-learn function known as "grid search"[3] is used for this purpose. It evaluates the performance of the algorithm in a cross-validation environment using different parameter values and combinations. The number of folds for CV was set to 10 and the evaluation was only performed on the training set. Finally, the optimal values for the algorithm parameters are obtained from the CV results and the model is saved in that setup for the testing phase.

For more information on the algorithms, their parameters, and grid-search please refer to the official Scikit-learn documentation[4].

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

[4] `https://scikit-learn.org/stable/modules/classes.html`

6.1.3 Training

After the setup, the models are trained on the training data. For most supervised[5] ML algorithms, the training can be expressed as an optimization problem. The objective is to find the optimal mapping function $f(x)$ to minimize a task-specific loss function on the training set,

$$\min_{\theta} \quad \frac{1}{n} \sum_{i=1}^{n} L(f(\vec{x_i}, \theta), y_i), \tag{6.1}$$

where n is the number of training samples, θ the parameter of the mapping function, L the loss function, $\vec{x_i}$ the feature vector of the ith sample, and y_i the associated label [201]. An example is *least squares* regression, where the goal is to minimize the sum of squares of the differences between true and predicted labels.

For this work, multiple ML models are trained on chronological age regression and majority classification based on the mentioned data and algorithms. They solve the tasks using the following combination of features:

1. Anthropometric measurements (AM)
2. Ossification stages of the three growth plates of the knee (OS)
3. Score of the knee joint (SKJ)
4. AM and SKJ

Given the feature combinations, the number of algorithms, and the two tasks, a total of 44 model variants are trained and evaluated. Refer to the Scikit-learn documentation[6] for more information on the specific algorithms, their parameters, and their optimization function.

6.2 Method 2: CNN-MRI

Neural networks are capable of learning and extracting information that is relevant to a specific task [187]. This assumes that the amount of samples they learn from and the complexity of the problem, enable them to find correlating features. With

[5] "Supervised learning refers to a class of systems and algorithms that determine a predictive model using data points with known outcomes" (definition by DeepAI)

[6] `https://scikit-learn.org/stable/modules/classes.html`

this in mind, the initial idea of this work was to use CNNs to solve the age estimation task based on the pre-processed (chapter 4) MRIs, without the need to extract age-relevant features via segmentation (chapter 5) or other techniques. Hence, an architecture similar to the contracting path of the CNN for segmentation (Fig. 5.8) was adapted to the age regression problem. It proved to be rather unsatisfactory to predict the age of a subject based on the pre-processed MRIs only. The training was unstable and early convergence for the validation loss was observed (Fig. 6.5). This suggests *underfitting*, i.e. the model is not capable of generalizing well on new data (*generalization gap*). Even after optimizing the hyperparameters of the aforementioned CNN and training during more epochs, a similar outcome was observed.

The conclusion from these initial observations is that the problem needs to be simplified. The following hypothesis is defined: by reducing the MRIs to the age-relevant structures via bone segmentation, a stable age estimation is possible.

This section describes how *Method 2* (Fig. 6.1) enables the simplification of the age regression problem in order to evaluate the hypothesis. The first part prepares the data for the CNN by extracting age-relevant objects from the MRIs (subsection 6.2.1). Next, a CNN architecture is engineered for the new scenario (subsection 6.2.2) and trained on *masked* 2D images (subsection 6.2.3). The model predic-

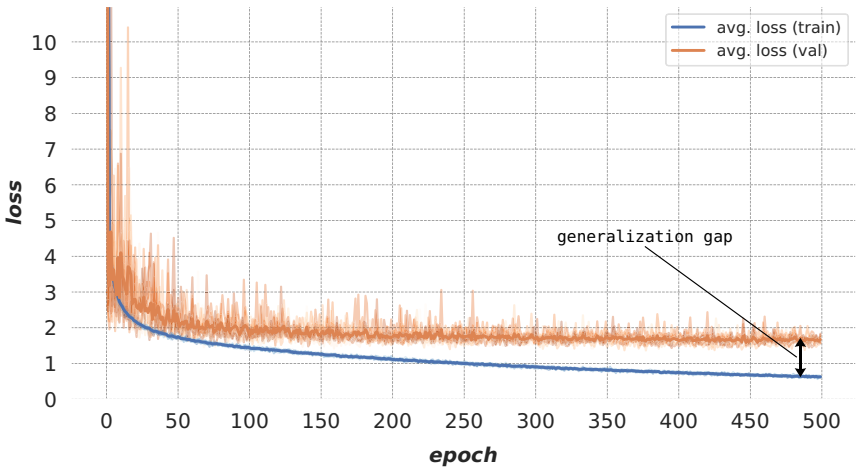

Figure 6.5: Train vs. validation losses for the age regression task using 2D MRIs *without* the bone segmentation step. The loss on the validation set converges early while the training set loss continues improving. This is a sign of underfitting (generalization gap).

tions per image slice are then sorted by subject and used to train ML algorithms on two tasks: the regression of the chronological age of a subject and on the majority classification using the 18-year-limit. The usage of ML algorithms enables the possibility of incorporating the AM and OS as additional features to solve the tasks.

Method 2 is applied to both coronal and sagittal MRIs of this work. For simplicity, the following subsections describe the method on the basis of the coronal images.

6.2.1 Data Preparation

Input data. The input data of *Method 2* are the preprocessed MRIs, the predicted segmentation maps, and the subject ages. The maps are acquired from the best segmentation model resulting from the 5-fold cross-validation (subsection 5.7). Afterwards, they are multiplied with the preprocessed MRIs to generate *masked images* (Fig. 6.6). The subject ages are retrieved from the formatted filenames of the images (Fig. 4.3) and therefore no further data has to be supplied.

Due to the slicing technique of MRI, the outer slice of an MRI volume can exhibited sparse tissue information. This holds for the masked images as well, which show limited or no bone structures in the outer slices. Supplying a CNN for age estimation with this sparse information can cause a misconception of how the actual age is deduced from the images. Especially, when training a model for age estimation based on 2D image data where context information from adjacent slices is missing.

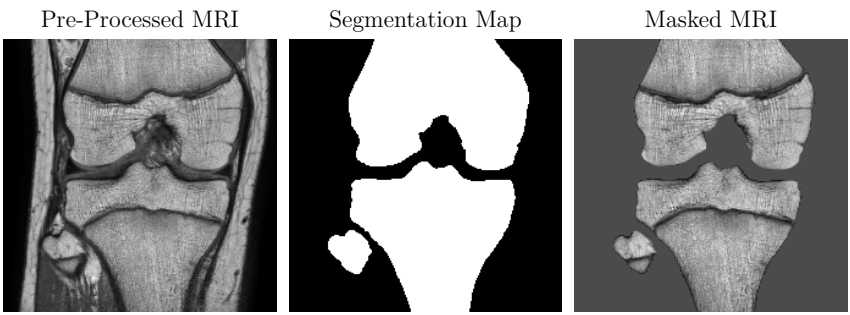

Figure 6.6: Pre-processed MRI slice (left), predicted segmentation map (middle), and masked image slice (right). Using the maps to mask the images enables the extraction of age-relevant structures in the image.

Removal of sparse bone information. The removal of sparse information in the images is attained in two data reduction phases (Fig. 6.7). The *first reduction phase* consists of removing all image slices containing little or no bone structures. This is achieved by first computing the total size of the segmented area of a 3D image and then removing all image slices with less than 2% of the total size. In the *second reduction phase*, the number of slices per masked image stack is reduced to a predefined minimum of 12 slices for all MRIs. This is necessary for several reasons. First, images from *Dataset A* and *Dataset A* have 41 and 24 slices, respectively, which causes an imbalance in the data used for the CNN. Second, it ensures that the slices of different subjects contain similar bone information. Third, it enables the possibility not only to train a neural network for age estimation based on 2D but also on 3D image data. CNNs using 3D input data require equal dimensions across all axes. Lastly, given that *Method 2* is based on 2D image data and predicts an age for each image slice, ML algorithms can be trained on the age predictions of all slices. During reduction phase 2 several actions are performed. At first, the amount of bone structures per slice of a 3D image is computed based on the segmentation map. Then, a reference image slice is defined as the starting point to extract the minimum of 12 slices. The reference slice is obtained from the bone-amount distribution per slice by calculating the center of gravity (COG) of the distribution. Finally, slices are evenly selected below and above the reference to match the minimum of 12 slices. The whole process of removal of sparse information is executed separately per stack of 2D masked images. The pseudo-code is shown in Algorithm 4.

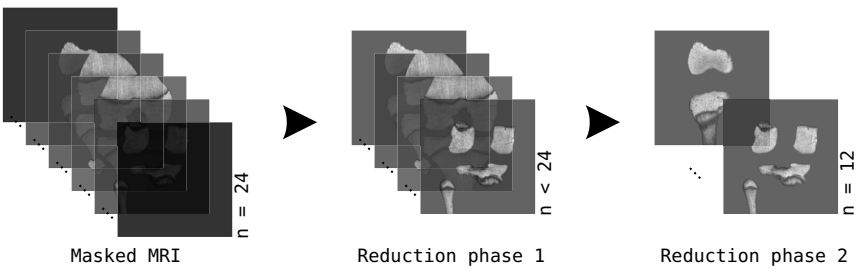

Figure 6.7: Removal of sparse bone information in two phases: first, the slices from the masked MRI that contain sparse or no bony structures are removed and subsequently, a predefined minimum of 12 slices are evenly extracted from the image volume for all MRIs. The reductions ensure a balancing of the CNN input.

Algorithm 4: Removal of sparse bone information

Input: Stack of masked knee MRIs
Output: Reduced stack to 12 slices
`// Phase 1: Remove image slices with sparse bone information`
for *each stack of masked images* **do**
  for *each image slice* **do**
    compute number of segmented pixels
  end
  compute total amount of segmented pixels
  remove image slices with less than 2% pixels of the total amount
end
`// Phase 2: Reduce stacks to 12 slices`
for *each reduced stack of masked images* **do**
  compute the center of gravity (COG) of the segmentation distribution
  select the slice at the COG as the reference
  extract 5-6 evenly spaced slices, starting from the reference, in ventral and dorsal direction to match the minimum of 12 slices
end

Dataset splits. One-hundred and eighty five coronal MRIs, each masked and reduced to 12 slices, are split into training, validation, and test sets. The split is performed *separately per dataset* (*Dataset A* and *Dataset B*) such that all sets have data from both. Moreover, the split for each source is done *separately per age group* to ensure that all sets have a similar age distribution. Also, it is assured that the training set includes the whole age range (14-21 years), such that the model can effectively predict any age in that range. Table 6.1 shows the exact splitting per source and age group, as well as the final number of MRIs per set.

Augmentation. The age distributions of both *Dataset A* and *Dataset B* are normally distributed (Fig. 3.2). While a normal distribution of values is generally not a problem, in ML one attempts to attain a balance between classes, in this case the age groups. Given a normal distribution the model might be incline to predict the mean class to improve it's performance. To mitigate the possible bias, a more uniform distribution in the training set is attained via 2D image augmentation. This

Table 6.1: Split per dataset and age group into three sets for the age estimation task based on coronal MRIs ($N = 185$)

Dataset	Age group (years)	n	Training set ($\approx 63\%$)	Validation set ($\approx 18\%$)	Test set ($\approx 19\%$)
A	14	7	5	1	1
	15	16	10	3	3
	16	20	12	4	4
	17	20	12	4	4
	18	16	10	3	3
	19	13	9	2	2
	20	8	6	1	1
	21	3	1	1	1
B	15	2	1	0	1
	16	19	13	3	3
	17	30	18	6	6
	18	16	10	3	3
	19	11	7	2	2
	20	4	2	1	1
Total	-	185	116	34	35

process is performed on the images of the *less represented age groups* and done *separately for each dataset* (Fig. 6.8). The final age distribution of the training data is not perfectly uniform, since the age range of subjects in each datasets is different. The image augmentation consists primarily of generating two additional FOVs of an image. If the maxmimum number of samples in an age group of the training set is not reached, further augmentation in form of translation and rotation is performed. The maximum number of modifications of a single sample is limited to 12:

- *FOV change* with a factor $\in \{0.9,1.0,1.1\}$
- *Horizontal translation* in pixel $\in \{-18,18\}$
- *Rotation* around z-axis in degrees $\in \{-5,5\}$

Finally, the number of 2D samples amounts to $n_{tr} = 2412$ for the training set, $n_{va} = 408$ for the validation set, and $n_{te} = 420$ for the test set. To enable the incorporation of AM and OS into the ML algorithms that follow the CNN, these features need to be duplicated times the number of image slices and augmentations for each subject to match the total number of 2D image samples.

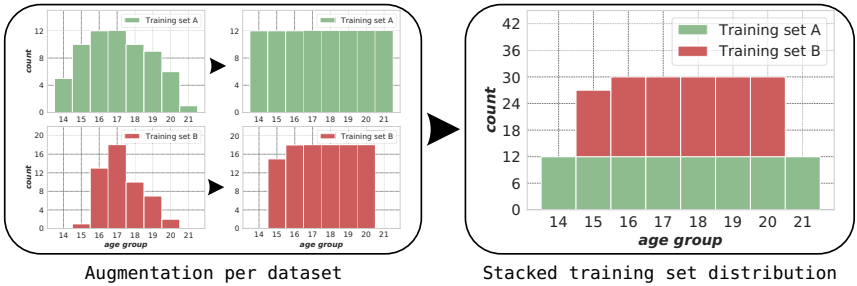

Figure 6.8: Augmentation of the training set is performed separately per dataset (left) and results in a more uniform distribution for the age estimation task (right)

6.2.2 CNN Architecture

The neural network for age estimation is created as a variant of the CNN of segmentation. Thereby, only the contracting path is adopted (Fig. 5.8) and adapted to the new problem. The network takes in 2D masked and reduced samples (subsection 6.2.1) with a size of 224×224 pixels and outputs the chronological age per sample in years (Fig. 6.9).

Due to the complexity of the age regression problem, several modifications to the architecture were necessary. In contrast to the CNN for segmentation, keeping a constant number of kernels per convolutional layer was not sufficient to find correlating features in the images. Therefore, the depth of the network is increased to five and the number of filters is doubled per depth starting with eight kernels in the first up to 256 kernels in the last convolutional layer. This modification improved the perfor-

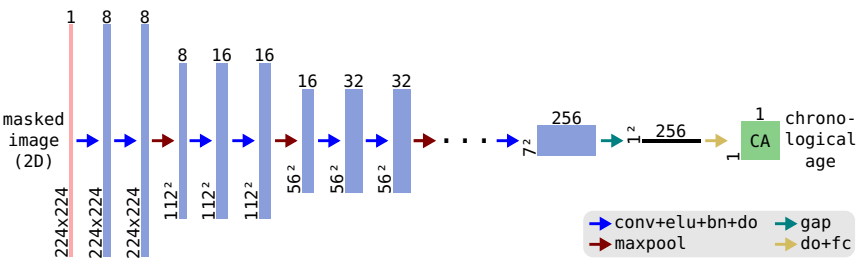

Figure 6.9: CNN architecture for age regression based on masked 2D knee MRIs

mance but resulted in high overfitting and unstable training. Batch-Normalization proved to be an optimal technique to stabilize training and was incorporated after each convolutional layer. To counter overfitting, Dropout is introduced after each convolutional layer, except at input level, using a hierarchical strategy. Dropout is set to 0.1 an incremented by 0.1 with each network depth. Similarly to the CNN for segmentation, ELU is set as activation function after convolutional layers and 2×2 Max pooling ($stride = 2$) is used to compress the images. At the deepest part of the network, the feature maps are reduced to a minimum size of 7×7 pixels.

To solve the regression problem, the final layers of the CNN have to be altered in comparison to the segmentation. Generally, one or more fully-connected (FC) layers, followed by an activation function, are added to the end of a network to transform the feature maps to the targeted output variable in regression or the output classes in classification. FC-layers connect all input nodes with output nodes, making them computationally expensive and prone to overfitting. This can be overcome, to some extent, by adding dropout layers [121]. Since all nodes are connected to the output, FC-layers act as black boxes, which makes them difficult to interpret and to identify the importance of each learned pattern in the image [121, 230]. One possible solution is to use a *global average pooling* (GAP) layer instead of the FC-layers [121, 230]. GAP computes the average of each feature map, reducing dimensionality from $h \times w \times d$ to $1 \times 1 \times d$, where h is the height, w the width, and d the depth of the maps. Since GAP only averages the feature maps, no weights need to be adjusted at this layer by the optimizer. Consequently, overfitting cannot originate at this layer contrary to an FC-layer, where a substantial amount of weights are optimized during the learning process. An alternative to GAP is *global max pooling* (GMP) which computes the maximum instead of the average of feature maps. In [230] the authors state that using GAP encourages the network to learn several important patterns in the image for the task, while GMP stimulates the model to identify the most crucial one.

Finally, the following layers are appended to the last convolution layer of the CNN for age estimation: GAP $\rightarrow$ Dropout ($p = 0.5$) $\rightarrow$ FC-layer $\rightarrow$ Linear activation function (Fig. 6.9). The FC-layer connects all 256 feature map to the output via a linear activation, which multiplies features with their corresponding weights to generate the estimated chronological age. The full architecture has a total of 1.18M trainable weights which are adjusted by the optimizer during training.

6.2.3 Training

The aforementioned CNN for age estimation was trained on coronal and sagittal MRIs, formerly pre-processed (chapter 4), segmented (chapter 5), and masked and reduced to 12 slices (subsection 6.2.1). Each image slice is considered as a separate sample during the training process.

Before the actual training, the trainable parameters of the network, i.e. the weights and biases, have to be initialized. When trained for the first time, the weights are initialized with random values, drawn from a normal or uniform distribution, and the biases with zeros. Alternatively, these parameters can be initialized with values learned from related but different problem. This "transfer of knowledge" is known as *transfer learning*. Since the segmentation task of this work also uses knee MRIs, the features learned on this image data can be reused for age estimation. Unfortunately, the CNN for age estimation (Fig. 6.9) differs in several layers to the contracting path of the segmentation network (Fig. 5.8). To enable transfer learning nonetheless, the CNN for age estimation is extended with an expanding path. This autoencoder is subsequently trained anew on the image data. Afterwards, the learned weights from the downsampling path of the autoencoder can be directly transferred to the CNN for age estimation.

Next, the hyperparameters to train the network for age estimation are defined and include: loss function, optimizer, learning rate, batch size, and epochs. Typical *loss functions* used for regression are the *mean squared error* (MSE), the *root mean squared error* (RMSE), or the MAE. Initial test runs showed, that training the CNN for age estimation using MSE, delivers lower errors and better handles outliers. MSE is defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{6.2}$$

where n are the number of samples, y_i the true age of the ith subject, and $\hat{y}_i$ the corresponding predicted age. Adam is selected as optimizer to minimize the loss function. The learning rate is reduced from 0.001 (section 5.5) to 0.0001, to fine-tune the initialized model for the age regression task. The total number of training samples ($n_{tr} = 2412$) are passed to the network in batches during the learning process. To find the best batch size for the task and underlying data, several test runs are performed with values ranging between 8 and 64. The optimal batch size of 16 resulted from these tests. Finally, the last hyperparameter to be set

is the number of training cycles, i.e. epochs. Due to the complexity of the problem and the low learning rate the number of epochs is set to 1000. Early stopping was not necessary since the final architecture did not show any signs of overfitting.

6.2.4 Age Regression

To attain a single and final age estimation per subject based on knee MRIs, the 12 age predictions by the previously defined CNN have to be combined. One possibility is to take either the average or the median of the predictions. Another option is to use the minimum-age concept (chapter 2). Unfortunately, these methods produced poor results. The alternative is to use ML algorithms similar to section 6.1. They can learn how important each of the 12 age predictions associated to one subject is, to predict the final chronological age. Moreover, ML algorithms facilitate the inclusion of a finite number of features for training. This enables the use of AM and SKJ (chapter 3) as additional features to the 12 age predictions.

For *Method 2*, the following combination of features is possible:

1. Age predictions from 2D coronal images
2. Age predictions from 2D coronal images, AM, and SKJ
3. Age predictions from 2D sagittal images

Finally, a total of 15 model variants are available for training given the three feature combinations and the five selected regression algorithms (subsection 6.1.2). AM and SKJ are not available for subjects with sagittal MRIs.

6.2.5 Majority Classification

The straightforward approach of *Method 2* to determine whether a subject is under or over 18 years, is to use the result from the regression models. Hence, if the regressor predicts an age below 18.00 years, than the individual can be classified as minor. An alternative approach is to use the 12 age predictions and the other data acquired from the subjects (AM and SKJ) and train a ML algorithm on classification *instead* of regression. Following an initial test phase, the latter proved to be the better solution for majority classification and was evaluated further (section 7.3).

Ultimately, a total of 18 model variants are evaluated and trained on the classification task given the six selected ML classifiers (subsection 6.1.2) and the same three combinations of data as used for the regression task.

6.3 Method 3: CNN-MIXED

Method 3 is the last approach for the age estimation of young males adults investigated in this work. The method is based on knee MRIs as well, similar to *Method 2* (section 6.2), and in addition, incorporates the numeric data from the subjects (AM and SKJ) into a new "multi-input and mixed data" CNN (Fig. 6.10). The data used by the multi-input CNN undergoes the data preparation (subsection 6.2.1) as well.

The CNN is designed as a neural network composed of two distinct branches, one for the image data and one for the combined numeric data acquired from the subjects. The *left branch* is a copy of the CNN from *Method 1* (Fig. 6.9) except for the last layer, which is replaced by a new FC-layer with five outputs. The *right branch* is a multi-layer perceptron. It has five input neurons to accept all the numeric data, i.e.

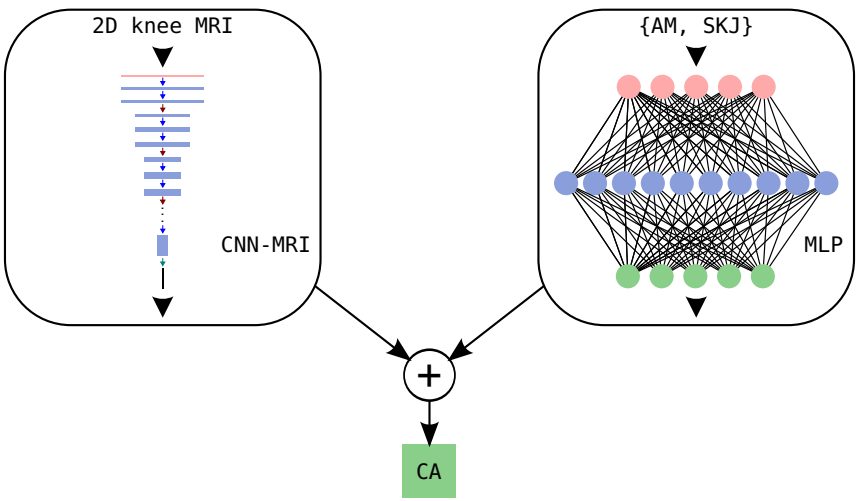

Figure 6.10: A "multi-input and mixed data" architecture to combine 2D knee MRIs and numeric data (anthropometric measurements (AM) and score of the knee joint (SKJ)) in one model

weight, standing height, sitting height, LLL, and SKJ. At the core, the MLP has a single fully-connected hidden layer with ten neurons. Each neuron takes in the five inputs and applies a non-linearity using an ELU activation function. The hidden layer is followed by Dropout ($p = 0.5$) to account for overfitting. The output layer is a FC-layer as well with five neurons. Here, each neuron receives ten inputs from the hidden layer and generates one output by applying a further ELU activation function.

The outputs of both branches are then concatenated and passed to a final FC-layer. This layer uses the *dense* representation of the image and numeric data learned from both branches to regress the chronological age. A linear activation is used for this purpose. Ideally, the equal number of outputs of both branches leads to a balanced weighting of image and numeric data for age estimation.

The hyperparameters for training, are similar to the ones from *Method 2* (subsection 6.2.3). The differences are in the learning rate, which is set to 0.0005, and the number of epochs, which are set to 100. In general, the optimization of the multi--input CNN proved to be challenging since both branches have to be adjusted by the optimizer. Convergence during training started early, which was a reason that the epochs were limited to 100.

After the training of the CNN, the same steps as in *Method 2* are performed for age regression and majority classification. The difference here is, that AM and SKJ can no longer be used to train the ML algorithms. Moreover, numeric data is only available for subject with an MRI examination in *coronal* slice orientation. As a consequence, the sagittal MRIs are disregarded for this method. Finally, there are just two trainable model variants for *Method 3*, one for regression and one for classification.

6.4 Model Evaluation

The performance of all models from *Method 1*, *Method 2*, and *Method 3* on age estimation is evaluated on the test set, i.e. the part of the data the model has never seen nor learned from. Furthermore, an unbiased estimate of a model performance is achieved with *stratified k-fold cross-validation*. For this work, $k = 5$ folds are generated by splitting the data into training, validation (only for *Method 2* and *Method 3*), and test set five times. Each time, the test set includes different subjects

in order to get independent evaluations. Additionally, to enable the comparison between the three age estimation methods, the test set of a respective fold contains the same subjects, independent of the method chosen.

To attain a higher reliability of the estimate of the model performance, each fold is evaluated a total of ten times. This is necessary due to the stochastic nature of deep learning and most ML models. Training a model several times on the same data will result in different prediction each time. The more robust the model, the lower the differences will be. The final evaluation is defined as "extended" stratified 5-fold cross-validation. The *stratification* ensures that the age distribution is approximately equal in all sets and all folds.

Age Regression

The following evaluation metrics are selected for the age regression task: MAE, *standard deviation of the absolute error*, *maximum absolute error*, and the *percentage of samples within 1-year and 2-years of absolute error* between the true and predicted chronological ages. The MAE is defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|(y_i - \hat{y}_i)\right|, \tag{6.3}$$

where y_i is the true chronological age of the ith subject, $\hat{y}_i$ the corresponding prediction by the model, and n the total number of subjects of the evaluation. As a reference for the existing variability, the values from a direct statistical evaluation of the training data (*stat*) are computed as: $\hat{y}_i = \bar{y}$, where $\bar{y}$ is the mean age of the samples in the training set. Thus, *stat* merely predicts the mean age for all subjects in the training set.

Majority classification

The evaluation metrics for the majority classification task are the *accuracy*, the *sensitivity*, and the *specificity* (18-year-limit):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6.4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{6.5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6.6}$$

All subjects below 18 years are defined as TPs. Hence, the accuracy represents the number of correct predictions (whether minor or adult) over the total samples of the test set. The sensitivity describes the number of minors that are correctly classified as such, while specificity expresses the number of adults that are correctly classified so. All three measures range in between 0 and 1, where values closer to 1 represent better performance.

The *Receiver Operating Characteristic* (ROC) curve is a further statistical analysis of the classification performance. In an ROC curve, the *true positive rate* (TPR) is plotted against the *false positive rate* (FPR) for various thresholds; TPR is just the sensitivity and FPR is 1 - specificity. These thresholds are used to decide between minors and adults based on the prediction of the classifier, which lies between 0 and 1. A higher threshold results in more minors being correctly classified while a lower threshold has the adverse effect of more adults being predicted correctly. Hence, the ROC analysis offers a possibility to find the optimal trade-off between sensitivity and specificity depending on the desired outcome for the underlying problem [111]. Finally, the *Area Under the Curve* (AUC) score measures the area under the ROC curve and indicates the capability of the model to distinguish between two classes (here minors and adults). The score ranges between 0 and 1 as well, 1 being a perfect classifier. An AUC score of 0.5 means that the model is classifying by chance.

Similar to the regression task, a statistical evaluation of the training data is used as a reference. For classification, *stat* predicts all subjects of the training set as being minors. Hence the reference metrics are a sensitivity of 1.0, a specificity of 0.0, and an AUC score of 0.5.

7 Results

7.1 Preprocessing Results

Bias Field Correction

Bias fields due to magnetic field inhomogeneities were observed in several of the available MRIs but were corrected to a great extent using the popular N4ITK [210, 211] algorithm (Fig. 7.1). The first column in the figure represents the original image slice, the second the corrected image, the third the color-coded bias field as overlay, and the last column the difference of the corrected and original image. The color-coded bias field highlights the under- (blue/purple) and overexposed (yellow/red) regions in the images which can appear in varying forms and locations in the image.

Case 1 shows underexposure on the bones, the structure of interest for this work, in contrast to higher intensities of the fat tissue. BFC effectively restores the balance (second image in top row). Case 2 depicts a large underexposed region at the top of the image which affects the shaft of the Femur (see first and third image). N4ITK corrects the intensities in the region to a great extent, making the femoral shaft and Patella more visible. Case 3 is a good example for intensity inhomogeneity within bone structures which is vastly restored with the algorithm. Case 4 shows large intensity differences between Tibia and Femur (red area in bias field). The difference is corrected with BFC, which enhances the grey values of the Femur and diminished the ones of the Tibia. Case 5 shows extensive underexposure of the bones, especially in the shaft region of the Tibia. Here, the BFC enhances tissue contrast in the image and partially restores the visibility of the tibial shaft.

In addition to intensity inhomogeneities, MRIs of this work were also affected by artefacts and BFC was only able to *partially* correct these images (Fig. 7.2). The N4ITK algorithm assumes a noise-free scenario and can thus not remove MR artefacts. Nevertheless, underexposed regions in the images are enhanced and the differentiation between different tissues improved. Case 1 shows a noisy and underexposed image. The subject had a large circumference of the subject's leg and therefore, no knee

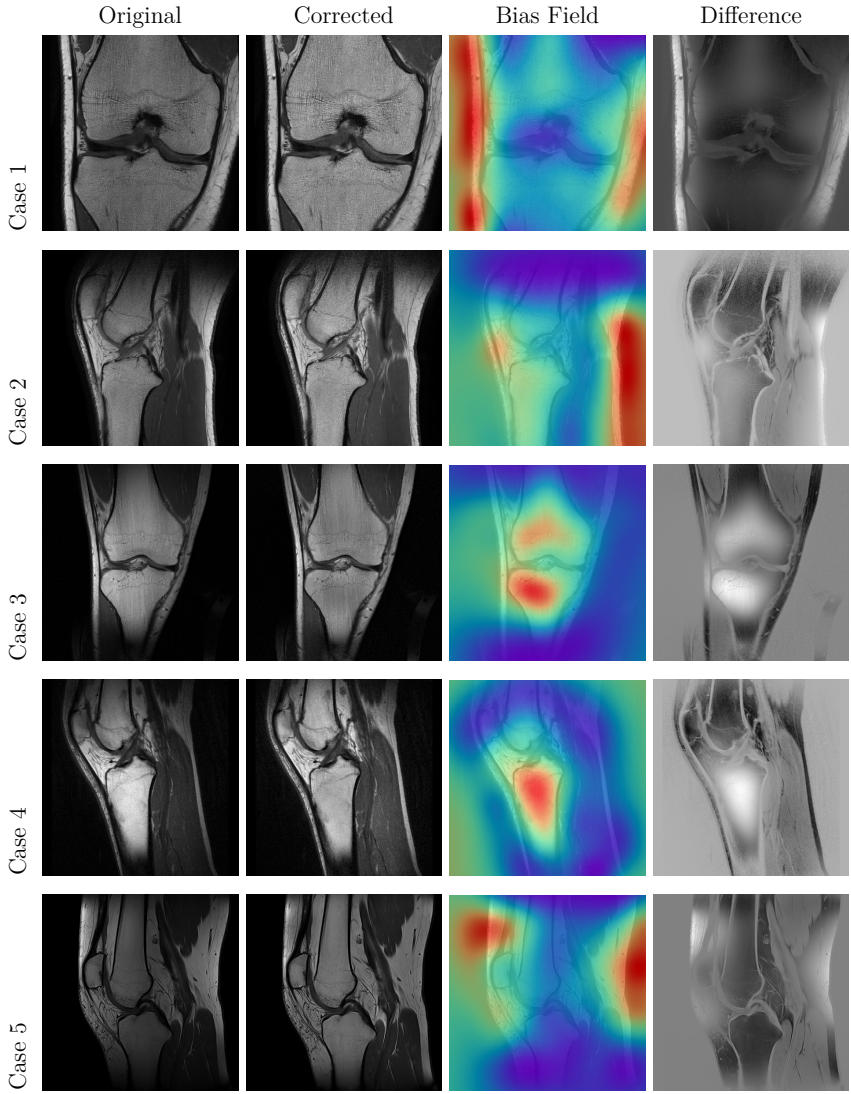

Figure 7.1: MRIs affected by intensity inhomogeneities were corrected using N4ITK [210, 211] algorithm. Original, corrected, color-coded estimated bias field as overlay, and difference of corrected and original image are shown from left to right for several cases. The color-coded bias field highlights under- (blue/purple) and overexposed (yellow/red) image regions representative of the inhomogeneities.

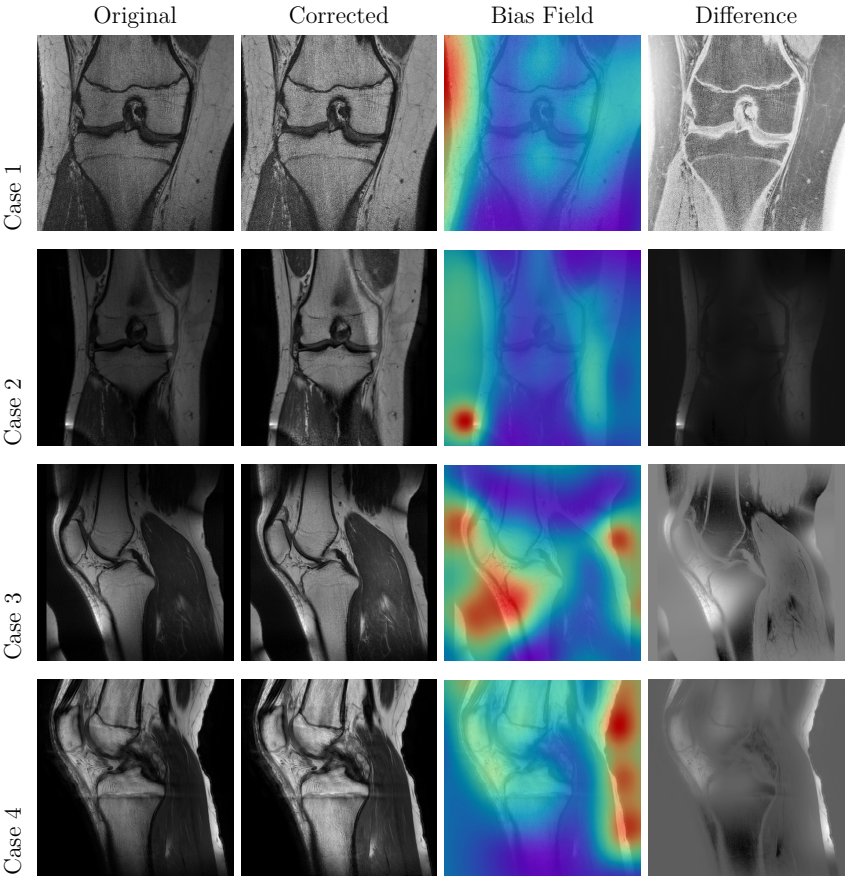

Figure 7.2: MRIs affected by intensity inhomogeneities *and MR artefacts*. The N4ITK [210, 211] algorithm assumes a noise-free scenario and can hence not remove image noise (Case 1), wrap-around (Case 2 and 3) or motion artefacts (Case 4). Nevertheless, the intensities are partially corrected to improve the homogeneity within a tissue.

coil could be used which led to the degraded quality of the MRI. BFC does not remove the noise in this case but is able to improve the contrast in the image. Case 2 depicts *wrap-around* artefacts, i.e. overlapping structures. In Case 3 there is a tiny circular region with extremely high intensities which causes the rest of the image to be underexposed. Additionally, a wrap-around artefact can be seen on the left side of the image overlapping with the Femur and other tissue. N4ITK cannot remove the wrap-around artefacts but is able to correct the induced inhomogeneities in the image to a great extent. Case 4 shows streaking artefacts due to motion which even affect the growth plate of the tibia. Again, BFC is not able to reverse this effect but improves tissue contrast due to field inhomogeneities.

The impact of different image sizes and spacings on the execution time of N4ITK was recorded (Table 7.1). Reducing an image from it's original size of $512 \times 512 \times 24$ voxels to $448 \times 448 \times 24$ voxels (approximately 25%) translated directly to the reduction of the algorithm duration by 25%. A decrease in image spacing, e.g. from 0.56×0.56 mm^2 to 0.45×0.45 mm^2, generally led to an increase of the execution time. The algorithm is dependent on the central processing unit (CPU) since it operates on multiple threads. This has to be taken into account given the reported values. The hardware details of the workstation available for this work can be found in Appendix A.

Table 7.1: Impact of image size and in-plane spacing on the execution times of the N4ITK [211] algorithm for bias field correction of 3D MR images

Orientation	State	Image size	Image spacing (mm^3)	Execution time (min:sec)
coronal	original	$512 \times 512 \times 24$	$0.39 \times 0.39 \times 3.9$	$\approx$ 06:01
coronal	resampled	$448 \times 448 \times 24$	$0.45 \times 0.45 \times 3.9$	$\approx$ 04:30
sagittal	original	$864 \times 864 \times 50$	$0.17 \times 0.17 \times 2.2$	$\approx$ 38:15
sagittal	resampled	$448 \times 448 \times 50$	$0.34 \times 0.34 \times 2.2$	$\approx$ 9:47
coronal	resampled	$448 \times 448 \times 24$	$0.45 \times 0.45 \times 3.3$	$\approx$ 04:00
coronal	resampled	$448 \times 448 \times 24$	$0.47 \times 0.47 \times 3.3$	$\approx$ 03:03
coronal	resampled	$448 \times 448 \times 24$	$0.49 \times 0.49 \times 3.3$	$\approx$ 02:26
coronal	resampled	$448 \times 448 \times 24$	$0.56 \times 0.56 \times 3.3$	$\approx$ 02:08

Automated Cropping

Automated cropping based on patch matching was used to extract a standardized VOI in each 3D MRI to compensate for differences in leg position of the subjects during the MR examination and for differences in FOV (section 4.3). Next, a few examples showing high variance in the position of the knee joint in coronal and in sagittal images (Fig. 7.3, Fig. 7.4). The characteristic region, or *patch*, is depicted in the first column, the original image in the second column, the resulting correlation map between the patch and the image in the third column, an the generated standardized VOI in the fourth column.

For coronal knee MRIs, the selected patch shows the interconyloid eminence (Fig. 7.3) and for sagittal MRIs, it is a different patch showing the posterior cruciate ligament (Fig. 7.4). In all the cases represented in the figures, the algorithm successfully detects the best fit of the corresponding patch in the image and is color-coded as red in the correlation map. Finally, the examples show, how automated cropping enables the extraction of a standardized VOI, irrespective of the position of the knee joint and the selected FOV during the MRI examination.

The execution time of automated cropping, i.e. patch matching followed by the extraction of the standardized VOIs, was tracked for images with different spacings, number of slices, and orientation (Table 7.2). The duration was similar for coronal and sagittal MRIs, although the latter had higher spacing. Analysing the orientations separately, the execution time was dependent on the spacing of the image. The lower the spacing, the higher the automated cropping duration.

Table 7.2: Execution times of the automated cropping step for two images with different spacing and size

Orientation	State	Image size	Image spacing (mm^3)	Execution time (sec)
coronal	downsampled	$448 \times 448 \times 24$	$0.45 \times 0.45 \times 3.3$	≈ 5.36
coronal	downsampled	$448 \times 448 \times 41$	$0.35 \times 0.35 \times 2.2$	≈ 8.39
sagittal	downsampled	$448 \times 448 \times 24$	$0.56 \times 0.56 \times 3.9$	≈ 5.30
sagittal	downsampled	$448 \times 448 \times 24$	$0.45 \times 0.45 \times 4.4$	≈ 8.46
sagittal	downsampled	$448 \times 448 \times 32$	$0.40 \times 0.40 \times 3$	≈ 9.95

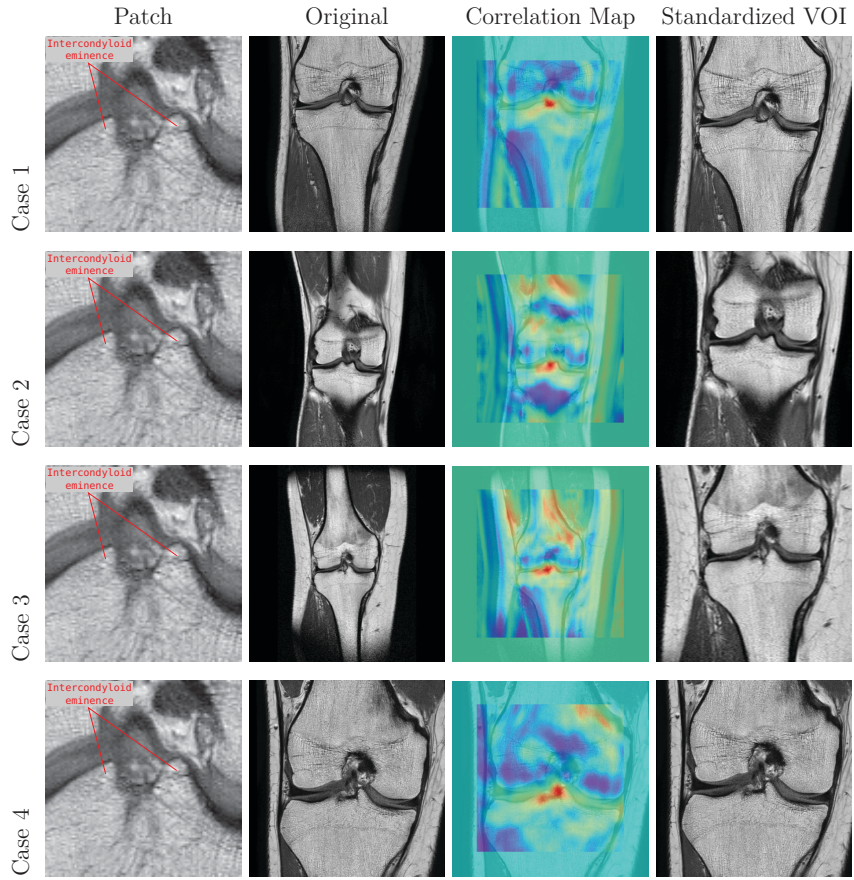

Figure 7.3: Extraction of standardized VOIs in coronal knee MRIs using a method based on patch matching. A patch showing the intercondyloid eminence is slid across the original image and the best position matches the red area in the correlation map. This point is the center of the resulting VOI. The automated cropping generates similar VOIs regardless of the position of the knee joint (Case 1 and 2) and independent of the FOV (Case 3 and 4).

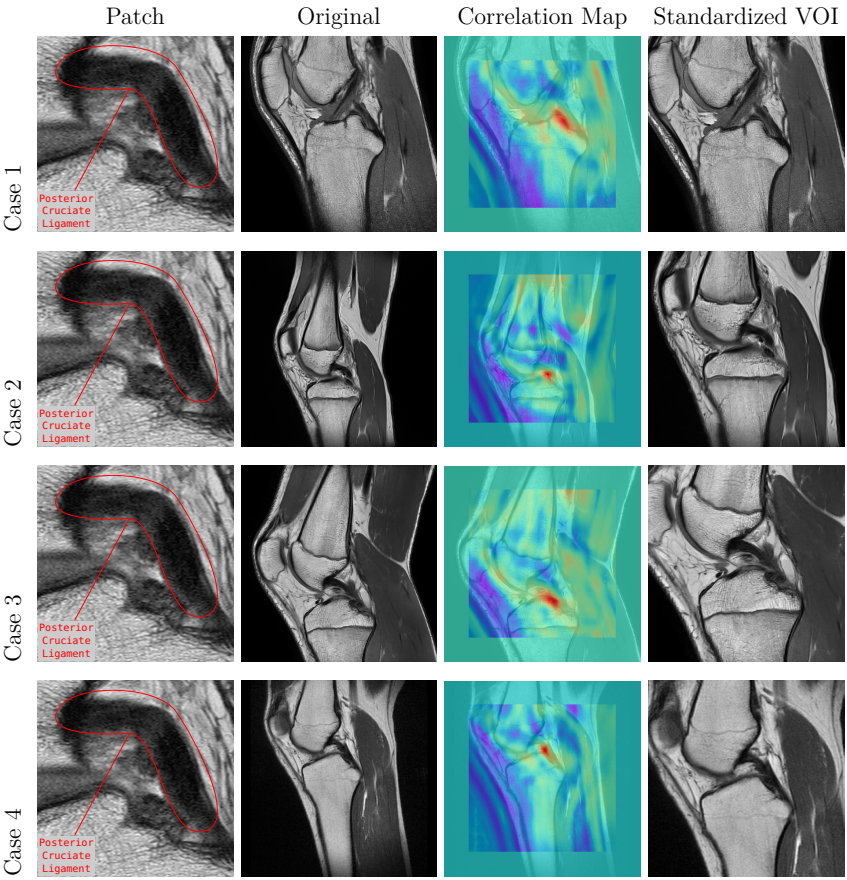

Figure 7.4: Extraction of standardized VOIs in sagittal knee MRIs using a method based on patch matching. A patch showing the posterior cruciate ligament is slid across the original image and the optimal position is color-coded as red in the correlation map. This point is the center of the resulting VOI. The automated cropping generates similar VOIs regardless of the FOV (Case 1 and 2) and independent of the position of the knee joint (Case 3 and 4).

7.2 Segmentation Results

Qualitative Results

Segmentation Quality

Before analyzing the quantitative results, it is often helpful to get an idea of the quality of the segmentation on the basis of visual and qualitative results (Figures 7.5 and 7.6). The figures show the MRI input to the *merged model* in the first column, the gold-standard segmentation or *ground truth* as overlay in the second one, the prediction of the model as overlay in the third one, and the difference between ground truth and prediction in the last column.

Highly accurate model predictions with DSC scores of over 98%, generally differed to the ground truth at the bone edges (Fig. 7.5). These differences can be caused by several factors, e.g. an imprecision of the manual labelling of the bones, prediction errors by the model, or also likely, the loss of information of the ground truth due to the downsampling step in the pre-processing (sections 4.2 and 5.3). Nevertheless, the quality of the predictions in these cases is profoundly high and hardly noticeable without the difference image.

In contrast, there were also cases with clear discrepancies between ground truth and prediction (Fig. 7.6). In case 1 the learning capacity of the model can be appreciated as it detects parts of the Tibia which were not labelled in the ground truth. The second case shows predictions that do not entirely cover the faintly visible part of the femoral condyle. Moreover, the model fails to separate Tibia from Fibula. In Case 3, the model appears to detect another structure on the right side of the image just under the Tibia. This could be a consequence of the horizontal flipping performed in the augmentation step (section 5.2) since the Fibula is likely to appear in that area of the image. Finally, in the last case the model is incapacitated by the wrap-around artefact on the right side of the image to fully detect the femoral condyle.

Model Exploration

The "black-box" effect of CNNs was mitigated by visualizing the learned feature maps the segmentation network (Fig. 7.8). The figure shows the intermediate sum of feature maps at different depths of the network. Along the contracting path,

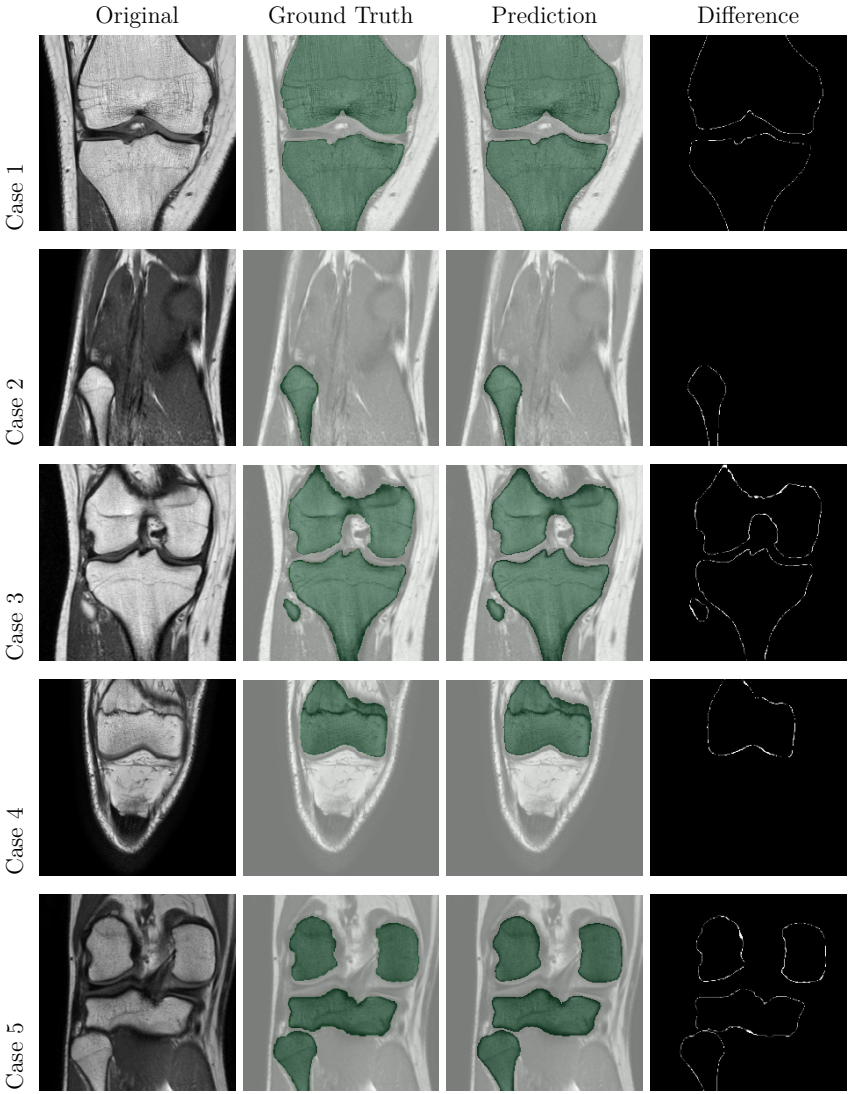

Figure 7.5: Excellent segmentation results with DSC scores of > 0.98. The figure shows the original MRI slice, the ground truth, the prediction by the model, and the difference between ground truth and prediction (from left to right) for several cases.

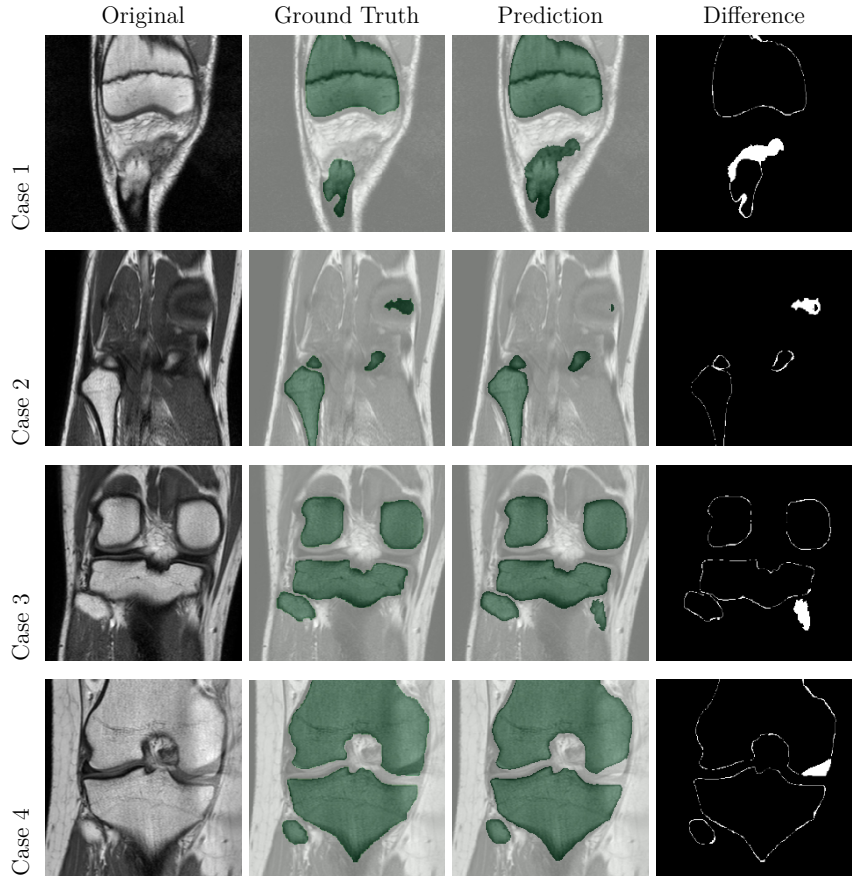

Figure 7.6: Discrepancies between predicted and ground truth segmentations. Case 1 highlights the learning capacity of the model as it detects parts of the Tibia which were not labelled in the ground truth. In the second case the model fails to fully identify the faintly visible part of the femur condyle and slightly over-segments the tibia. In Case 3 the model falsely predicts tissue in the lower right part of the image as bone. This could be a result of the horizontal flipping performed during augmentation. The last case shwos how MR artefacts can have an impact on the model performance. A part of the femoral condyle is not detected due to the prominent edge caused by a wrap-around artefact.

the network learns low-level features such as edges (especially Down-Blocks 1 - 3). Afterwards, more complex structures can be seen but are more difficult to interpret (Down-Block 4 and Bottom-Block). On the expanding path, the CNN appears to form larger and more specific structures, possibly soft-tissue in Up-Blocks 1 and 2, until it finally begins to learn how to distinguish bone from other tissue (Up-Block 4). Finally, the feature maps of the last block are combined to predict the output segmentation.

A further visualization method to attempt to understand the learned features in more depth is called "activation maximization" [55]. Erhan et al. state that the goal of the method is to generate better qualitative interpretations of the features, especially of the high-level ones. The principle is to search for patterns in the input data which maximize the activation of a specific hidden unit, i.e. kernel. This is done via *gradient ascent* in the input space, i.e. the derivative of a filter activation is computed and the input sample is moved in the direction of this gradient.

Kernels from the downsampling path focus on the detection of low-level features such as edges in horizontal, vertical, and diagonal orientation (Fig. 7.8). The upper row in the figure depicts individual feature maps. The maximization of the corresponding kernel activations can be seen in the bottom row and confirms the previous assumptions that the kernels learn linear patterns in the image.

High-level features learn complex representation of the data and can be hard to interpret (Fig. 7.9). The upper row in the figure depicts individual feature maps which appear to be specialized in more complex structures. The leftmost filter could have learned parallel diagonal line pairs. The corresponding activation maximization below shows diagonal patterns which are more pronounced and sparse than the ones in Fig. 7.8. This shows how high-level features become specialized on specific structures and patterns in the image. The second column depicts a feature map which could represent a bone probability map at low spatial dimension. The maximization shows patterns agglomerated in the center of the image but is difficult to doubtlessly interpret. In the third column, the filter seems to have detected soft tissues surrounding the bones. The activation maximization shows a border-like pattern which could represent these complex arrangement of structures. The final example is a highly interesting discovery, as the a filter of the CNN appears to have learned a *growth-plate detector*. Related to the segmentation task, such a filter is reasonable since the network has to learn that the growth plates have to be included in the output. These have much lower gray-scale values than the bone and appear

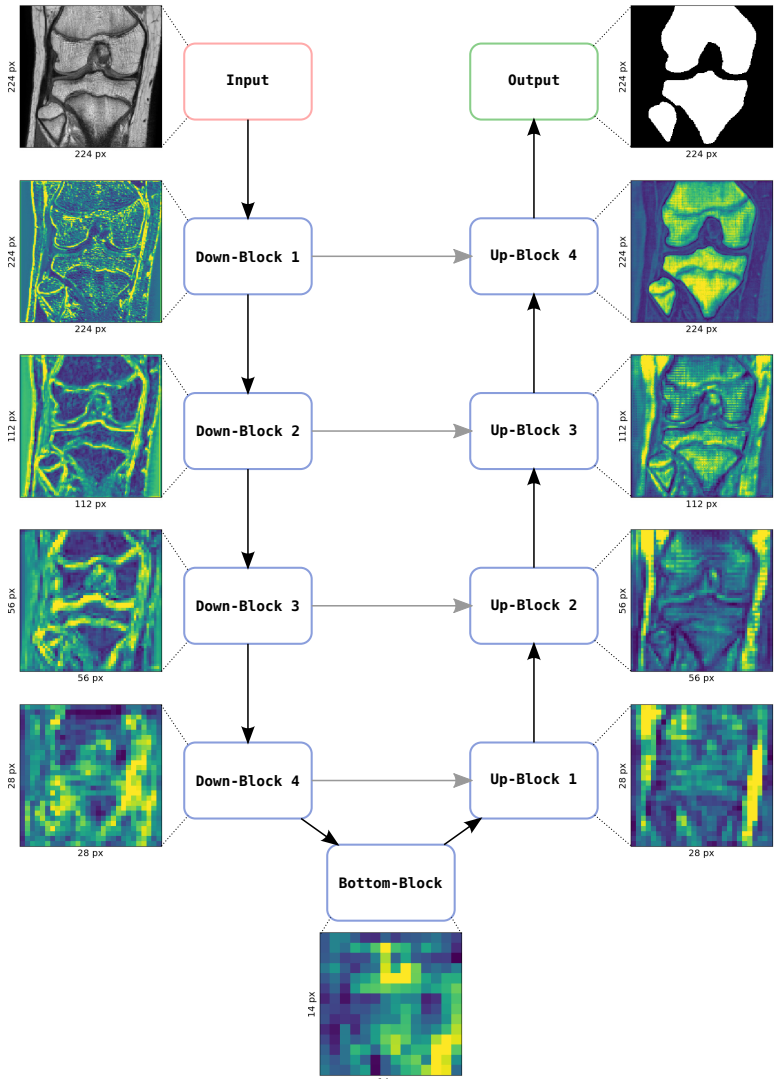

Figure 7.7: Intermediate sum of feature maps after each building block of the segmenatation network. The feature maps are normalized and color-coded for visualization purposes. In the first three Down-Blocks the network seems to focus on low-level features such as edges. The network then appears to be specializing on certain structures such as soft tissues (Up-Block 1 and 2). Finally, in Up-Block 4 most features are bone detectors.

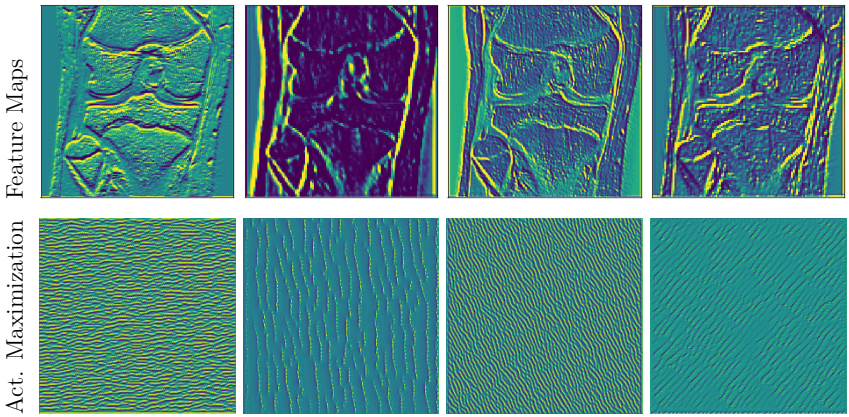

Figure 7.8: Visualization of individual feature maps of the segmentation network (top) and activation maximization of the corresponding filters (bottom). The CNN appears to learn edges, i.e. low level features, in horizontal (1st column), vertical (2nd column), and diagonal orientation (3rd and 4th column).

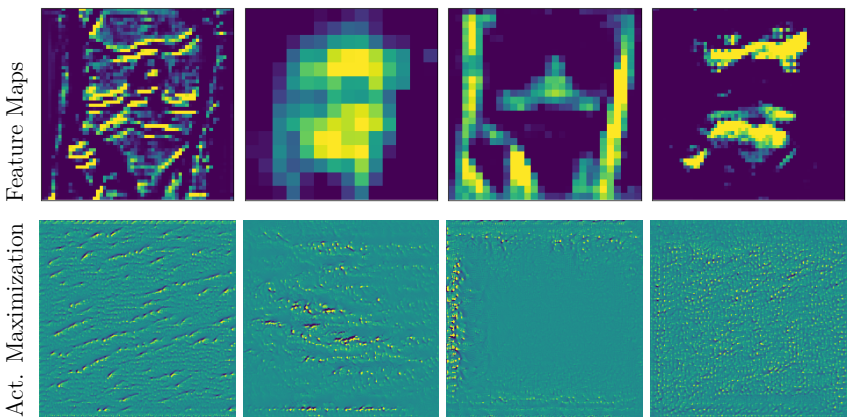

Figure 7.9: Visualization of individual feature maps of the segmentation network (top) and activation maximization of the corresponding filters (bottom). The CNN appears to learn more complex and high level features. The leftmost example resembles a double diagonal feature detector, the second one a bone probability map, the third one a soft--tissue detector, and the last possibly a growth-plate detector.

dark in the MRIs. Nonetheless, the CNN to has to learn to distinguish between the growth plates and other structures having low intensities as well.

Noise Exploration

The robustness of the segmentation architecture was evaluated on noisy input data. The noise was applied in form of dilation and erosion with a kernel size of 7 to the ground truth of the training and validation data. The effect of this noise on the test set was analyzed and visualized (Fig. 7.10).

To compare the model predictions to the ground truth, the same dilation or erosion from the training was applied to the sample of the test set (Fig. 7.10 — third column in uneven rows). From the qualitative results in the figure, one can deduce the model has learned to predict bone structures in between an over- and under-segmented area caused by the noise. The difference images to the right show how the errors diminish in comparison to the introduced noise.

Transfer Application and Learning

The merged model trained on coronal MRIs was used in two applications to segment data with a different extent and orientation (section 5.7). For simplicity, this model is referred to as the *coronal model*.

First, the aforementioned model was applied onto *uncropped and larger* coronal MRIs (Fig. 7.11), i.e. with a size of 448×448 pixels instead of 224×224 pixels as used in training. Qualitatively, the model performs well on detecting the bone in the uncropped image which has a larger FOV then the data the model was trained on. From the examples in the two central columns of Fig. 7.11, one can appreciate that the model does not fully detect the femoral shaft. These results can be expected since the FOV in the training data was smaller and hence, with less amount of visible bone shafts.

Second, was the application of the coronal model onto *uncropped and larger sagittal* MRIs (Fig. 7.12), data never presented to the model during training. The predictions are quite accurate, even for slices with little bone information (Fig. 7.12 — P1). Even the Patella is detected, a bone the model was never thought to segment. However,

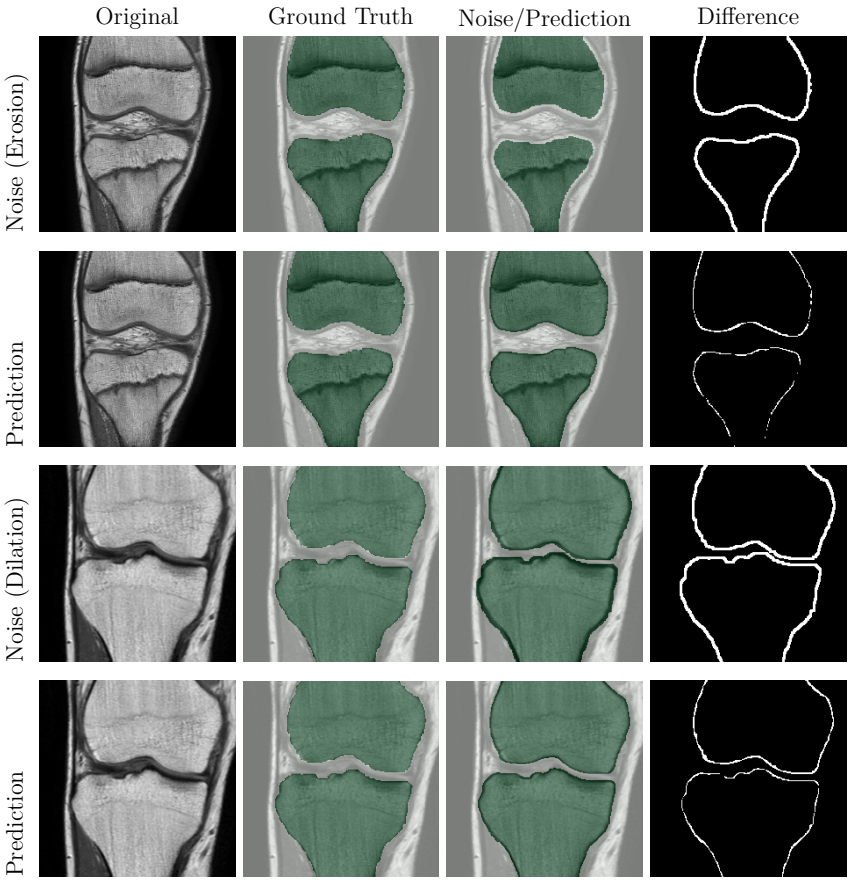

Figure 7.10: Segmentation quality of a model trained on noisy data. The figure shows the original MRI slice, the ground truth, the applied noise or the prediction, and the difference between ground truth and prediction from left to right. The uneven rows present the dilation or erosion applied to the ground truth of a sample from the test set. The even rows show the corresponding bone predictions of the model trained on noisy training data.

the model has problems fully detecting the shafts of the bones since the training data never included such large FOVs.

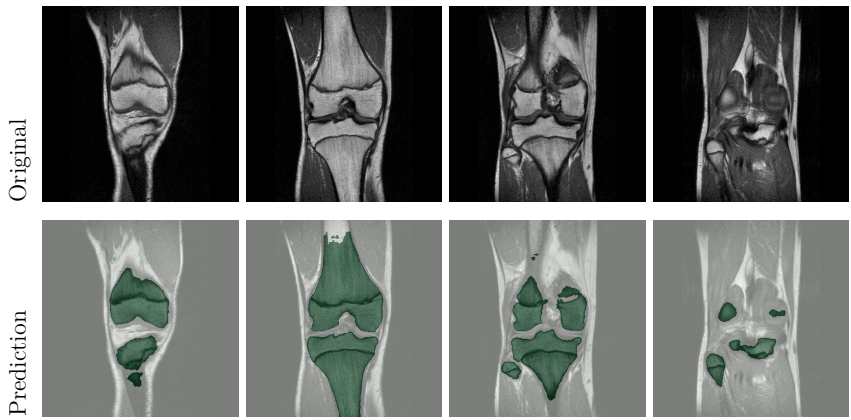

Figure 7.11: *Unknown and uncropped* MRI slices (top row) were segmented using a model trained on standardized coronal images. The model performs well on detecting most of the bone structures in images it has never learned from, but fails to fully detect the bone shafts due to a much smaller FOV present in the training data (bottom row).

The results from the second application were improved by retraining the merged model on ground truth data of sagittal MRIs (subsection 5.7). The improvements induced by the *transfer learning* approach are plainly visible (Fig. 7.12 — P2, Δ (P1-P2)). The fine-tuned model is able to fully segment the bone structures in sagittal images (P2) that were only partially recognized by the merged model trained on coronal MRIs (P1). Moreover, the fine-tuned model learned to correctly ignored the Patella. This bone was explicitly *disregarded* in the ground truth samples used for the retraining since it was not considered a bone relevant for age estimation. The merged model was not "taught" to ignore the Patella and it is a rather appealing result that the model detects the bone altogether. The fine-tuned model was additionally able to overcome the insufficient detection of the bone shafts by the merged model (third column) to a great extent.

These qualitative results of the transfer learning approach show the potential of CNNs even when provided with less training data (25 ground truths samples for sagittal versus 100 for coronal samples). Related to the age estimation task, the

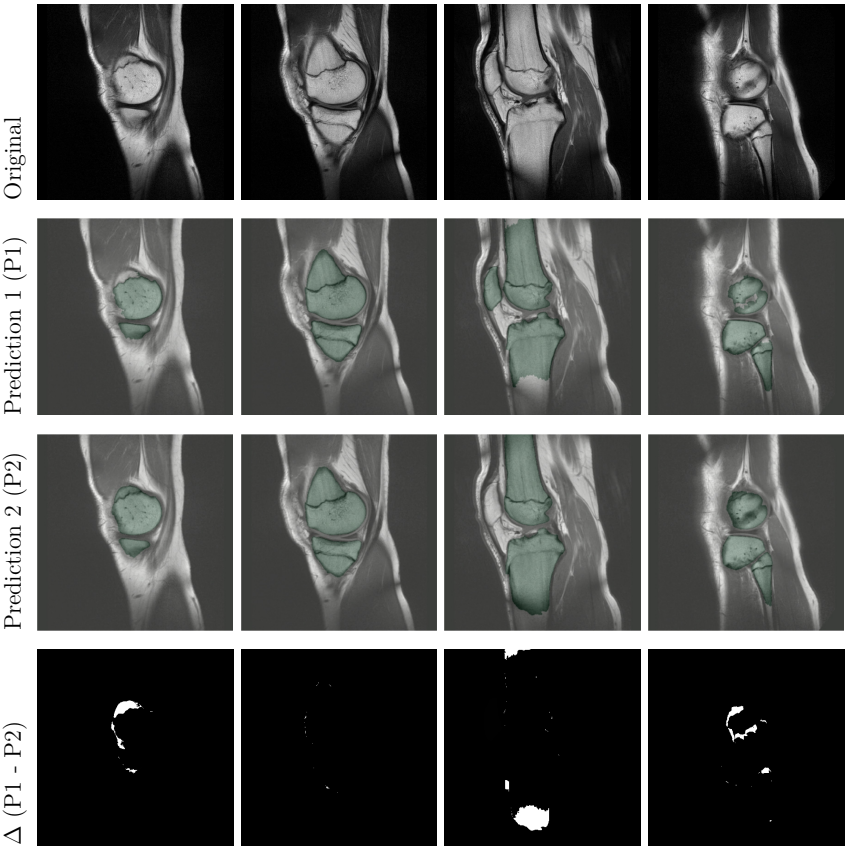

Figure 7.12: Model predictions on unknown and uncropped MRI slices from a different orientation using the merged coronal model (P1) and the fine-tuned model trained on sagittal ground truth data (P2). The predictions of the merged model (P2) are relatively good considering that it has never seen images in sagittal orientation. It even detects the Patella, a bone which was never labelled in the coronal training data. The enhancement through the fine-tune model is visible in the third and fourth row. It detects more bone structure than the merged model and ignores the Patella (as it was taught to). The fine-tuned model still has problems detecting the entire bone shafts which is expected since a smaller FOV was present in the training data.

need for fine-tuning becomes clear from the third and fourth columns of Fig. 7.12. Here, the mere application of the merged model on the differently-oriented sagittal images showed inaccurate predictions of bone structure close to the growth plates.

Model Performance

The loss of the CNN for segmentation (chapter 5) on the training and validation sets was tracked for five training rounds of one fold (Fig. 7.13). The progression of the losses is stable and does not vary much between training rounds. The losses are much higher during the early epochs, especially at the first one, due to the random initialization of the model weights before training. They rapidly decline after just one further epoch and then slowly head towards a plateau. The losses on the validation sets follow the same progression of the losses of the training sets and do not diverge. This suggests that the model generalizes well on new data. Moreover, no sign of overfitting is present. The losses curves reach values between 0.013 and 0.014 at the final epochs, which can directly be associated to DSC scores $\geq$ 98% for the training and validation sets and represent an excellent overlap between prediction and ground truth.

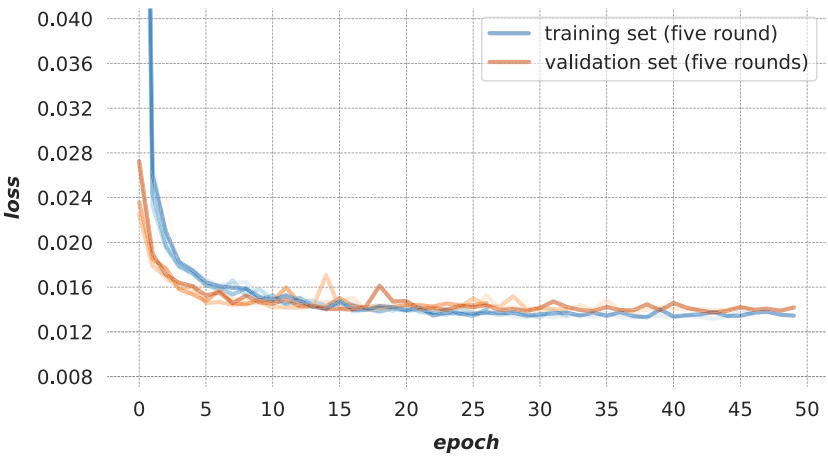

Figure 7.13: Training and validation loss for the merged model

Given the previous analysis of the loss progression during training, it is apparent that the designed network architecture was suitable for the segmentation task. Next,

the performance of several segmentation models are reported. These models were evaluated over multiple training rounds (Tab. 7.3). The first part of the table shows models evaluated using 5-fold cross-validation (subsection 5.7). The second part are the results using LOOCV (subsection 5.7) on the sagittal MRIs. The last part of the table represents the "experimental" results on noise exploration (subsection 5.7).

Table 7.3: Performance of various models on the segmentation task averaged over multiple training rounds (metrics in %)

Model type	DSC	IoU	Precision	Recall	Error
Merged *	98.5 ± 0.1	97.0 ± 0.1	98.5 ± 0.1	98.5 ± 0.1	0.7 ± 0.0
Femur *	98.6 ± 0.0	97.3 ± 0.1	98.6 ± 0.1	98.6 ± 0.1	0.3 ± 0.0
Tibia *	98.4 ± 0.1	96.8 ± 0.2	98.3 ± 0.4	98.5 ± 0.2	0.3 ± 0.0
Fibula *	96.2 ± 0.7	92.7 ± 1.2	96.8 ± 0.6	95.6 ± 1.6	0.1 ± 0.0
Combined *	97.3 ± 0.1	94.7 ± 0.2	99.6 ± 0.0	95.1 ± 0.2	1.3 ± 0.1
3D CNN *	98.3 ± 0.1	96.6 ± 0.2	98.2 ± 0.2	98.3 ± 0.2	0.8 ± 0.0
Sagittal +	97.4 ± 0.8	95.0 ± 1.4	98.3 ± 1.2	96.6 ± 1.7	1.1 ± 0.3
Noise #	95.2 ± 0.3	90.8 ± 0.6	91.6 ± 0.6	99.1 ± 0.1	2.2 ± 0.2

∗: evaluated on 5-fold cross-validation and repeated five times
+: evaluated on 5-fold cross-validation and repeated five times
#: evaluated on one fold and repeated five times

The overall best results for the test sets were acquired with the *merged model* which considered all knee bones as the target structure. The average DSC is 98.5% and the average IoU is 97.0%. Precision and recall are well-balanced with average values above 98%. The total error is below 1%. The standard deviations of all evaluation metrics for the 25 training rounds are below 0.2% suggesting that the designed architecture produced robust models for the segmentation task and the knee MRI data.

The *Femur and Tibia model* had comparable results to the merged one, the former even with slightly better evaluation metrics and lower standard deviations between training rounds. The *Fibula model* was the worst of the single bone model achieving 96.2% for DSC and an IoU of 92.7%. Nevertheless, the Fibula is the smallest bone, only occupying a small portion of the image frame and only available in relatively few slices of a 3D knee MRI. Therefore, lower results can be expected and still represent a very accurate concordance with the ground truth.

The *combined model* uses the predictions of the three separate bone models to create a segmentation of all bone structures in the image. It achieves a DSC of 97.3% and an IoU of 94.7% on average, which is inferior to the merged model. The precision with 99.6% is highest among all models while the recall with 95.1% is the lowest. The combined model has a larger total prediction error of 1.29% since the errors of the separate bone models aggregate.

The CNN based on 3D convolutions (see 5.7) achieves comparable results to the merged and combined models on the segmentation of all bone structures. All five metrics are only worse on a small scale.

The LOOCV on the model trained on sagittal MRIs resulted in an average DSC score of 97.4% ± 0.8%, an IoU of 95.0% ± 1.4%, a precision of 98.3% ± 1.2%, a recall of 96.6% ± 1.7%, and a total error of 1.1% ± 0.3%. The results are comparable to the ones of the other models and are even promising since the sagittal model was only trained on 25 gold-standard segmentations.

Finally, the experimental results using noisy data can be extracted from the last section of the table. The average the DSC over five rounds on the test set is 95.2% (± 0.3%) which is noticeably lower compared to the other models. It indicates that noise in the ground truth clearly degrades the performance of the model. Nevertheless, the error of the model is lower than the error introduced through the noise (6.69%), which results in a DSC of 93.31% for the test set. This means that the predictions contain approximately 28% less noise than the data the model was trained on. The proposed network architecture for the segmentation is hence robust against noise in the training data to a certain extent. Thus, small errors during the manual segmentation by the expert should not have a large impact on the performance of the model.

In summary, the merged model, which segments all bones in knee images, achieves the best segmentation result with a DSC of 98.5% evaluated on basis of a 5-fold cross-validation. Refer to Appendix D for additional results on segmentation.

7.3 Age Estimation Results

Qualitative Results

Similar to the previous section, qualitative results are presented in advance. This is helpful to follow along with the decisions that were made during the development phase, especially for *Method 2*, and what impact they had on the performance of the models. Several aspects of the development phase are exemplified with the aid of results and are the following: (a) the benefit of using segmentation *prior* to age estimation based on knee MRIs, (b) the importance of reducing the images to a certain number of slices, (c) the need for ML algorithms in addition to CNNs for age regression, and (d) the advantage of training ML-based classifiers on the age predictions per image slice of the CNN.

The first aspect concerning the development phase of *Method 2* is the unsatisfactory training of the CNN based on unsegmented images and was already introduced in section 6.2. These result are now compared to the CNN trained on *masked* images (Fig. 7.14). The training, validation, and test sets used for both models contained the same images except for the processing step applied to them. This made a direct comparison possible.

The differences between both learning processes are readily visible. The model trained on unsegmented MRIs exhibits a rather fast convergence of the validation loss while the training loss improves steadily. This suggests that the model is starting to learn the training samples "by heart" and is not able to generalize on new data. The training process was limited to 500 epochs for this model since no substantial gains could be attained with longer training. Starting already at around 100 epochs, the loss curves of training and validation sets start to diverge. On the contrary, the model trained on masked and reduced images shows a more ideal training progression. The losses of both sets develop at a similar rate during training and are even slightly lower for the validation set over all epochs. These results support the hypothesis from section 6.2 that "by reducing the MRIs to the age-relevant structures via bone segmentation, a stable age estimation is possible".

The second aspect about the development phase of *Method 2* is the reduction of the image slices per MRI volume after masking them with the segmentation maps. The uncertainty here is, how much effect the content in the slice, i.e. the segmented

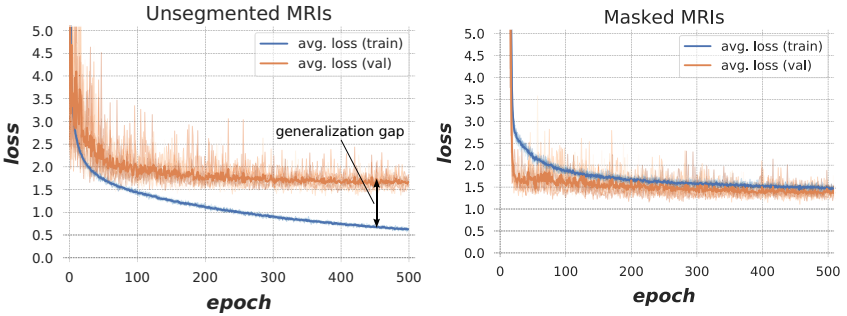

Figure 7.14: Training vs. validation loss (ten training rounds) for the age estimation task using two different CNN models based on *Method 2*: trained on unsegmented MRIs (left) and on masked and reduced MRIs (right). Training with unsegmented images leads to a generalization gap, while the use of masked MRIs improves the loss progression and delivers better results on age regression.

bone structures, have on the age predictions and if the removal of slices with sparse bone information was the correct approach.

Plotting the absolute error (AE) of the age prediction for each of the 12 slices of the reduced images shows, that lower (1-3) and higher (11,12) slices exhibit slightly higher median values, larger error bars, and a broader interquartile range (IQR) in comparison to central slices (Fig. 7.15). The boxplots in the figure show the median absolute age deviation as a line (orange), the box extending from the lower to upper quartile, the whiskers from $\pm$ 1.5 IQR, and the outliers as circles. The green line represents the average bone ratio per slice.

This results supports to some extent the decision to remove slices with even sparser bone information (section 6.2). It is a stimulating finding and could be considered in future analyses to verify, if selecting fewer slices of a 3D knee MRI and more centrally located in the volume, leads to lower age regression errors.

A third aspect of the development phase of *Method 2* is about the benefit of training an ML algorithm on the age predictions per slice by the CNN, to regress the final age of a subject. A graph is generated to compare the absolute prediction errors of the CNN, before and after applying an ML algorithm, visually per age group (Fig. 7.16). The figure shows boxplots with the same properties as in the previous figure, and a green line for the ratio of each age group's size to the training size. The reduction of the AEs through the ETR is clearly visible to the right of Fig. 7.16, as both the

boxes get narrower and the whiskers and outliers reduce in comparison to the left. For both cases, the predictions for the lower (14 and 15 years) and higher (20 and 21 years) age groups are more inaccurate. A feasible reason for this behaviour could be that these group were less represented in the coronal training data (green line).

Averaging or performing a minimum-age concept on the 12 age predictions per subject by the CNN, instead of using ML algorithms, led to higher errors. For the test sets and training rounds used in Fig. 7.16, the MAEs were 0.92 ± 0.82 years, 0.70 ± 0.54 years, and 0.69 ± 0.47 years for the minimum-age concept, "averaging", and the ETR. The basis were the predictions by the CNN which deviated 0.79 ± 0.62 years on averaged from the true chronological ages. These results confirm, that training an ML algorithm on the age predictions per image slice by the CNN, is the best approach to minimize the age regression problem.

Finally, the last aspect of the development phase was a on the classification task. Similar to the regression, ML algorithms were trained on the predictions of the

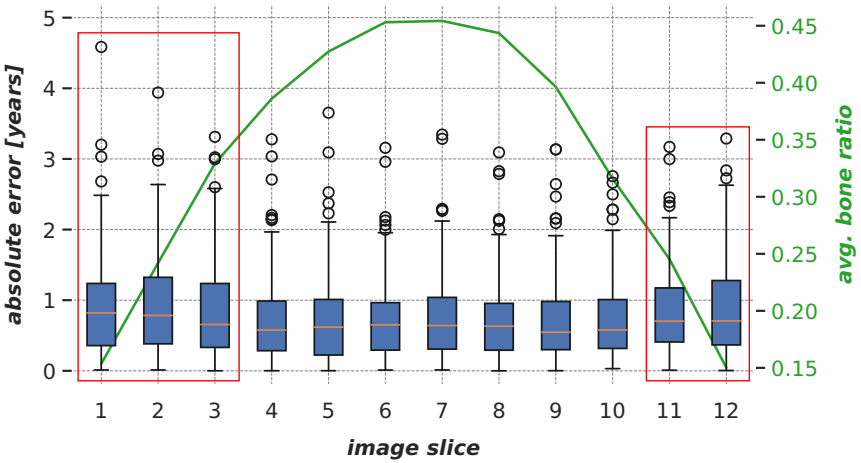

Figure 7.15: Absolute error between the true and predicted ages *per image slice*. Results are based on a CNN of *Method 2* for several training rounds. Lower and higher slices show slightly larger absolute errors than central ones (framed in red). This suggest that the higher bone content (green line) in the central slices positively contributes to lower errors in age regression.

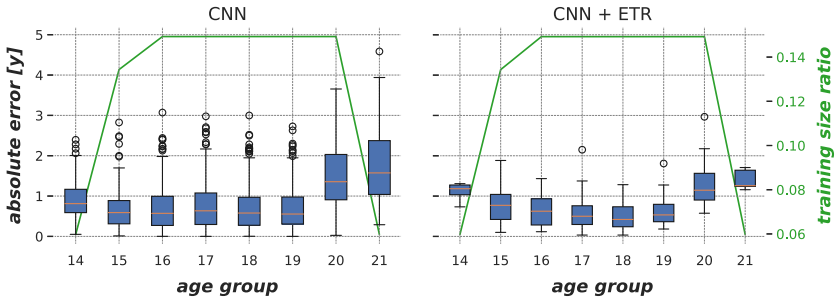

Figure 7.16: Absolute error between predicted and actual chronological age per age group for several training rounds and folds. The age prediction errors by the CNN from *Method 2* (left) are greatly reduced after applying the final regression using an ETR (right), i.e. a tree-based regressor. The higher error for age groups 14, 15, and 16 years could be associated to their corresponding ratio in the training data of the model.

CNN to solve the task. To understand this decision, consider the following two cases, where subjects were accurately classified using a trained ML-based classifier.

The first case is a subject aged 17.33 years who was correctly predicted a minor using a CNN followed by a classifier. The CNN age predictions for the 12 image slices ranged between 17.6 and 18.4 years (Fig. 7.17 — Case 1). The final predicted age by the regressor that followed the CNN, was 18.03 years and thus overestimated the subject's age by 0.7 years. In this case, the advantage of using a classifier on the CNN predictions can be appreciated since it correctly identified the minor.

The second case is a subject aged 18.08 years who was correctly predicted an adult using a training classifier. The CNN predictions for all slices ranged between 17.6 and 18.4 years (Fig. 7.17 — Case 2). The final age predicted by the trained regressor was 17.92 years, thus slightly underestimating the subject's age. Similar to the previous case, the predicted age by the regressor has the consequence that the individual is falsely classified, while the classifier makes the correct prediction.

Additional and appealing findings using the established method of "growth plate ossification assessment based on expert evaluation" used in practice, can be found in Appendix E.

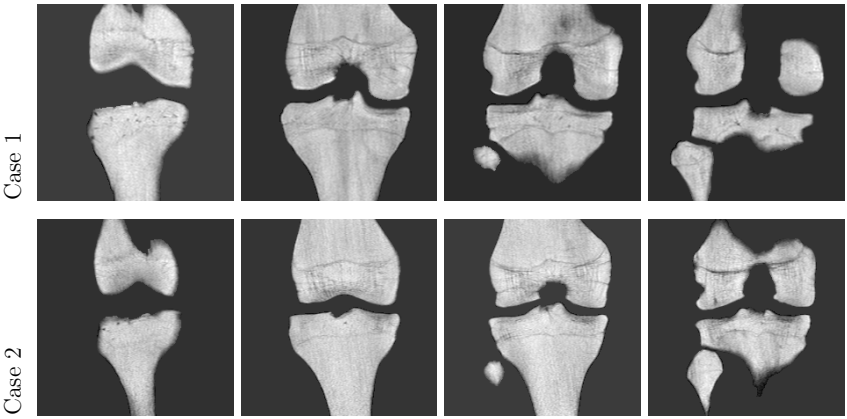

Figure 7.17: Several image slices of two subjects, one aged 17.3 years (Case 1) and the other 18.08 years (Case 2), showing almost full ossification. In both cases, a tree-based regressor deviates from the actual age, 18.03 for Case 1 and 17.92 for Case 2, which results in a false classification. On the contrary, a classifier is able to correctly classify the subjects in both cases.

Model Performance on Regression

The performance of multiple model variants using *Method 1* (section 6.1), *Method 2* (section 6.2), and *Method 3* (section 6.3) was analyzed for the age regression task. The models were evaluated on all ten training rounds (*all*) or only the best one of each fold (*best*). All regression results were compared to a direct statistical evaluation of the training data, designated as *stat* (subsection 6.4). The *stat* differed between methods since the number of samples and the distribution of the training data was different. However, the same subjects were included in the test sets of each fold, irrespective of the method. This allowed an unbiased comparison between *Method 1*, *Method 2*, and *Method 3*. The principal metric to define the performance of the models on age regression was the MAE. The performance was measured on the test set only, i.e. the part of the data *unknown* to the models. The test sets included $n_{cor} = 35$ subjects for all models of M1, M2 (coronal MRIs), and M3 and $n_{sag} = 75$ subjects for the models of M2 (sagittal MRIs).

To present the results in the tables, the following designations are necessary due to the high number of model variations:

- Method 1 (M1), Method 2 (M2), Method 3 (M3)
- Anthropometric measurements (AM), ossification stages of the knee growth plates (OS), score of the knee joint (SKJ), coronal knee MRIs (COR), sagittal knee MRIs (SAG)
- Support-vector regressor (SVR), extremely randomized trees regressor (ETR), gradient tree boosting regressor (GBR)

The results for *Method 1* show how ML algorithms trained on AM, OS, and SKJ improve the statistical evaluation of the training set (Table 7.4). For both *all* and *best* rounds, the lowest MAE was achieved with a combination of AM and SKJ, suggesting that more and different data can help to reduce errors on age regression. Using SKJ or OS was indifferent, but the former was preferred as it is a single feature instead of three. Finally, the best average MAE of 0.77 ± 0.60 years is attained with a support vector regressor. Less than 5% of test subjects from all folds had a deviation from the true age above two years.

Table 7.4: *Age regression performance* of several model variants from *Method 1* (M1) on the test sets in an "extended" 5-fold cross-validation using AM, OS, and SKJ

Rounds	Data	Regressor	MAE ± SD	Max AE	% ≤ |2.0y|
-	-	stat*	1.23 ± 0.90	3.93	79.20
all	AM	GBR	1.00 ± 0.70	2.63	90.93
	OS	GBR	0.91 ± 0.68	3.01	92.00
	SKJ	ETR	0.91 ± 0.68	3.01	92.00
	AM+SKJ	GBR	0.84 ± 0.63	2.62	94.86
best	AM	SVR	1.00 ± 0.69	2.57	92.00
	OS	GBR	0.90 ± 0.68	3.00	92.57
	SKJ	GBR	0.90 ± 0.68	2.98	92.57
	AM+SKJ	SVR	0.77 ± 0.60	2.58	95.43

∗: predicts all subjects with the mean age of the training set
all/*best*: all ten or best training rounds per fold are included

The *stat* of the models based on *Method 2* (MAE of 1.63 ± 0.99 years) was higher in comparison to M1 due to a larger and augmented training set. All model variants of M2 achieved better results than the *stat* (Table 7.5). Considering *all* training rounds, the MAE improved about 12% and the SD about 15% by using a regressor on the age predictions of the CNN. The inclusion of AM and SKJ as features to the ML regressors did not markedly improved the results. Overall, the model with the

best average performance on age regression, achieved an MAE of 0.69 ± 0.47 years and maximum AE of 2.15 years. It combined all available data and used CNN and ETR in succession to regress the final age of the subjects of the test sets.

Table 7.5: *Age regression performance* of several model variants from *Method 2* (M2) on the test sets in an "extended" 5-fold cross-validation using *coronal* knee MRIs, AM, and SKJ

Rounds	Data	Regressor	MAE ± SD	Max AE	% ≤ |2.0y|
-	-	stat*	1.63 ± 0.99	3.59	59.70
all	COR	CNN	0.81 ± 0.65	3.55	94.00
	COR	CNN+ETR	0.71 ± 0.55	2.46	96.78
	COR+AM+SKJ	CNN+SVR	0.73 ± 0.55	2.39	97.60
best	COR	CNN	0.79 ± 0.62	3.49	95.05
	COR	CNN+ETR	0.67 ± 0.49	2.10	98.86
	COR+AM+SKJ	CNN+ETR	0.69 ± 0.47	2.15	98.29

∗: predicts all subjects with the mean age of the training set
all/*best*: all ten or best training rounds per fold are included

Method 2 was applied to a larger dataset containing sagittal MRIs as well which altered the *stat*, in comparison to the coronal case, to an MAE of 1.93 ± 1.20 years and a maximum AE of 4.74 years due to the broader and more uniformly distributed age range of the training set. Similar to M1, the models of M2 were superior to *stat* and the errors substantially reduced with the inclusion of ML-based regressors to combine CNN age predictions (Table 7.6). The best performing model variant of M2 using sagittal MRIs attained an MAE of 0.79 ± 0.57 years and a maximum AE of 2.63 years. Less than 5% of predictions deviated more than two years from the actual chronological.

The final evaluated method on regression was M3, which integrated the AM and OS directly into the CNN, instead of including these features into the training of the ML algorithms following the CNN. Similarly to M1 and M2, the statistical mean of the training set (*stat*) was surpassed by the models based on M3 (Table 7.7). Likewise, the combination of CNN and ML regressors improved the results further. Ultimately, the lowest prediction errors were achieved by using the CNN on mixed data and an ETR in succession, attaining an average MAE of 0.71 ± 0.54 years over all five folds and maximum AE of 2.2 years.

Table 7.6: *Age regression performance* of several model variants from *Method 2* (M2) on the test sets in an "extended" 5-fold cross-validation using *sagittal* knee MRIs

Rounds	Data	Regressor	MAE ± SD	Max AE	% ≤ |2.0y|
-	-	stat*	1.93 ± 1.20	4.74	54.98
all	SAG	CNN	0.92 ± 0.73	4.31	90.91
	SAG	CNN+SVR	0.81 ± 0.62	2.86	94.85
best	SAG	CNN	0.89 ± 0.70	4.22	92.44
	SAG	CNN+ETR	0.79 ± 0.57	2.63	95.73

∗: predicts all subjects with the mean age of the training set
all/*best*: all ten or best training rounds per fold are included

Table 7.7: *Age regression performance* of several model variants from *Method 3* (M3) on the test sets in an "extended" 5-fold cross-validation using *coronal* knee MRIs, AM, and SKJ

Rounds	Data	Regressor	MAE ± SD	Max AE	% ≤ |2.0y|
-	-	stat*	1.63 ± 0.99	3.59	59.70
all	COR+AM+SKJ	CNN	0.92 ± 0.70	3.47	91.49
	COR+AM+SKJ	CNN+ETR	0.84 ± 0.65	2.74	92.78
best	COR+AM+SKJ	CNN	0.85 ± 0.64	3.29	94.33
	COR+AM+SKJ	CNN+ETR	0.71 ± 0.54	2.20	95.43

∗: predicts all subjects with the mean age of the training set
all/*best*: all ten or best training rounds per fold are included

Model Performance on Classification

All three age estimation methods of this work (M1, M2, and M3), were also evaluated on the majority classification (18-year-limit) with an "extended" 5-fold cross-validation. This included the evaluation of the models on all ten training rounds (*all*) or only on the best one of each fold (*best*). The reference statistical evaluation on the training set is designated as *stat* in the tables and represents a naive classifier with 100% sensitivity and 0% specificity, i.e. it applies the principle of "in dubio pro reo".

The designations of the model variants from regression are extended to include the ML algorithms for classification: k-nearest neighbours classifier (KNC), support-vector classifier (SVC), decision tree classifier (DTC), random forests classifier (RFC), extremely randomized trees classifier (ETC), gradient tree boosting classifier (GBC).

The age distribution of the subjects analyzed with *Method 1* was slightly imbalanced towards minors and thus resulted in an accuracy of 61.33% for the statistical evaluation of the training set. The weight of the classes was included into the ML algorithm before training to account for the imbalance. All fitted classifiers had a higher performance than *stat*, but did not surpass 80% in all metrics (Table 7.8). Learning from AM alone, delivered insufficient results. Using OS instead of SKJ as a feature for the growth plate maturation, gave slightly better metrics. The combination of AM and SKJ, did not improve the results as in the regression task, but rather hurt the accuracy and sensitivity. Finally, the best model for M1 was GBC based on OS as input data with an average accuracy of 81.14%, sensitivity of 82.73%, specificity of 78.46%, and AUC of 83.18%.

Table 7.8: Performance on *majority classification* of several model variants from *Method 1* (M1) on the test sets in an "extended" 5-fold cross-validation using AM, OS, and SKJ

Rounds	Data	Classifier	Acc.	Sens.	Spec.	AUC
-	-	stat*	61.33	100.00	0.00	50.00
all	AM	KNC	70.29	76.36	60.00	73.81
	OS	GBC	80.57	81.82	78.46	83.36
	SKJ	GBC	80.00	80.91	78.46	83.15
	AM+SKJ	ETC	74.29	69.09	83.08	83.92
best	AM	KNC	77.71	80.00	73.85	76.92
	OS	GBC	81.14	82.73	78.46	83.18
	SKJ	GBC	80.00	80.91	78.46	83.15
	AM+SKJ	ETC	76.74	71.55	85.54	85.87

∗: predicts all subjects in the training set as minors
all/*best*: all ten or best training rounds per fold are included

All listed classifiers of *Model 2* achieved above 80% in accuracy, sensitivity, specificity, and AUC (Table 7.9). The best performing classifiers on coronal MRIs were RFCs and incorporated either only the coronal MRIs or all data. Both classifiers surpassed 89% in accuracy. The RFC on MRIs only, had a slightly higher average sensitivity and AUC which could prove to be advantageous compared to the RFC on all data, which has a higher specificity, depending on the preferred outcome.

Method 2 was also trained on a larger number of sagittal MRIs with a distribution of the training set marginally inclined to minors (52.26%). The models performed better than *stat* when evaluated on *all* ten training rounds and on the *best* one per

Table 7.9: Performance on *majority classification* of several model variants from *Method 2* (M2) on the test sets in an "extended" 5-fold cross-validation using *coronal* MRIs, AM, and SKJ

Rounds	Data	Classifier	Acc.	Sens.	Spec.	AUC
-	-	stat*	49.25	100.00	0.00	50.00
all	COR	CNN+SVC	85.71	86.36	84.62	90.82
	COR+AM+SKJ	CNN+RFC	83.49	81.36	87.08	89.55
best	COR	CNN+RFC	89.14	89.09	89.23	92.52
	COR+AM+SKJ	CNN+RFC	89.71	88.18	92.31	91.99

∗: predicts all subjects in the training set as minors
all/*best*: all ten or best training rounds per fold are included

fold (Table 7.10). RFC trained only on sagittal MRIs attained the best average metrics, with an accuracy of 90.9%, a sensitivity of 88.6%, a specificity of 94.2%, and a AUC of 94.4% over all folds.

Table 7.10: Performance on *majority classification* of several model variants from *Method 2* (M2) on the test sets in an "extended" 5-fold cross-validation using *sagittal* MRIs

Rounds	Data	Classifier	Acc.	Sens.	Spec.	AUC
-	-	stat*	52.26	100.00	0.00	50.00
all	SAG	CNN+SVC	87.47	88.41	86.13	94.33
best	SAG	CNN+RFC	90.93	88.64	94.19	94.38

∗: predicts all subjects in the training set as minors
all/*best*: all ten or best training rounds per fold are included

The last method evaluated on majority classification of the 18-year-limit was *Method 3*. The gradient boosting classifier, i.e. GBC, achieved metrics under 80% — except for the AUC — considering all ten training rounds (Table 7.11). In contrast, the RFC was the best model with an average accuracy of 86.86%, a sensitivity of 85.46%, a specificity of 89.23%, and a AUC of 88.53% over all folds.

Table 7.11: Performance on *majority classification* of several model variants from *Method 3* (M3) on the test sets in an "extended" 5-fold cross-validation using *coronal* MRIs, AM, and SKJ

Rounds	Data	Classifier	Acc.	Sens.	Spec.	AUC
-	-	stat*	49.25	100.00	0.00	50.00
all	COR+AM+SKJ	CNN+GBC	76.34	74.91	78.77	83.96
best	COR+AM+SKJ	CNN+RFC	86.86	85.46	89.23	88.53

∗: predicts all subjects in the training set as minors
all/*best*: all ten or best training rounds per fold are included

Summary

In summary, several successful models were trained for the age regression and majority classification tasks. Comparing the three methods of this work (M1, M2, and M3), *Method 2* proved to be the optimal approach to solve both tasks.

For regression, M2 was best configured using a CNN trained on coronal MRIs followed by an extremely randomized trees regressor. The ETR used the CNN age predictions per image slice, the AM, and SKJ to regress the final chronological age of an individual. It achieved an average MAE of 0.69 ± 0.47 years and maximum AE of 2.15 years on the tests sets of five different folds. Each fold included 35 test subjects amounting to a total of 175 different subjects evaluated with the aforementioned model. The predictions of ETR are plotted over the true chronological ages of all test subjects (Fig. 7.18). The green central line highlights a perfect prediction, while the two parallel grey lines encompass 95% of the model predictions. The predictions lie relatively close and evenly distributed along the green line except a few "outliers" outside the area between the grey lines.

For classification, the best method was M2 as well but using sagittal instead of coronal MRI. RFC proved to be the best ML algorithm to learn from age predictions of the CNN to discriminate between adults and minors. It achieved a high performance on this task with an accuracy of 90.9%, a sensitivity of 88.6%, a specificity of 94.2%, and a AUC of 94.4% over all folds. The ROC curve suggests, that the model has the potential to increase its sensitivity at the cost of specificity, or conversely (Fig. 7.19).

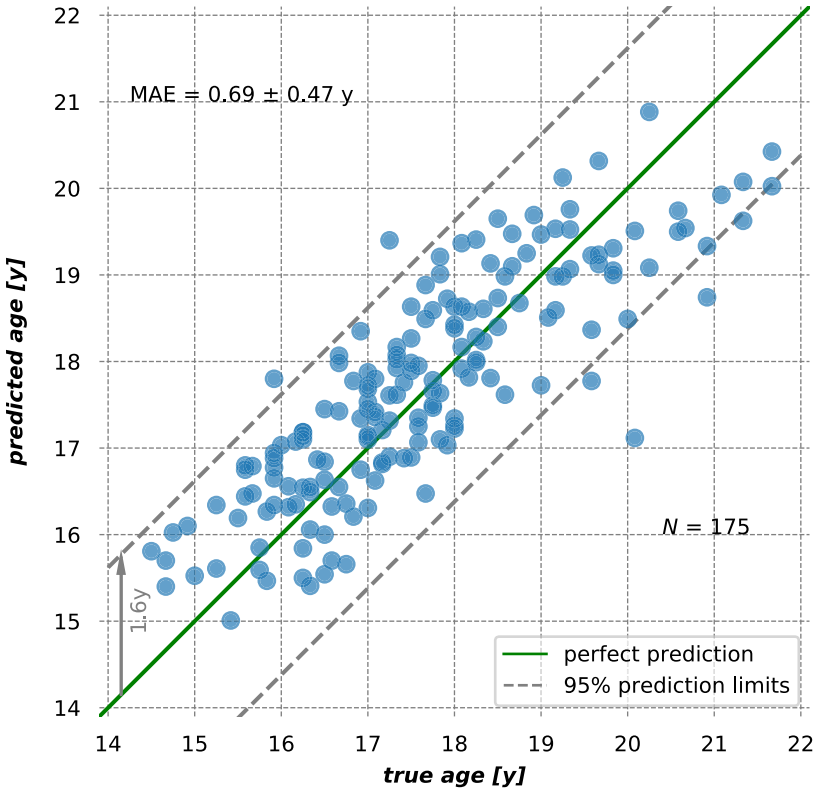

Figure 7.18: Predicted vs. true chronological age of test subjects from all five folds ($n = 35 * 5 = 175$) using a CNN followed by an ETR based on *Method 2*. The green central line highlights a perfect prediction, while the two parallel grey lines encompass 95% of the data.

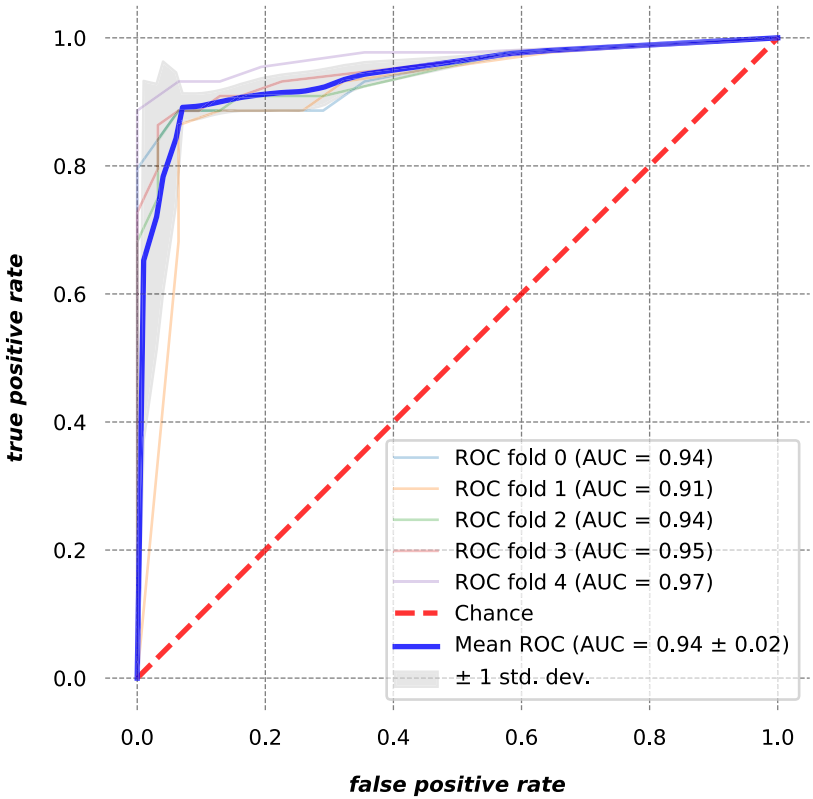

Figure 7.19: ROC curve for the best model on majority classification. The random forest classifier (RFC) attains an accuracy of 90.9%, a sensitivity of 88.6%, a specificity of 94.2%, and a AUC of 94.4% averaged over five distinct folds. The high mean AUC suggests, that shifting the threshold can further improve the sensitivity at the expense of the specificity, or vice versa.

8 Discussion

Population and Materials

The target population of this work was Caucasian male individuals between 13 and 21 years of age with a middle to high socio-economic status. The careful selection of the study participants created a particularly *homogeneous* population, which makes comparisons between studies more feasible since factors influencing growth differences between subjects are mitigated to a large extent.

In this work, the age estimation was performed on male individuals only, since gender has perhaps the largest influence on skeletal maturity (chapter 3). In other studies, e.g. [35, 52, 104], the growth plate ossification was analyzed on both sexes which can induce a high variability and unreliable results. EASO [57] reassures that gender has to be taken into account since methods show different margins of error for both sexes. Tscholl et al. [209] observed this behaviour when estimating the age of young female football players for Under-16 and Under-17 tournaments from MRI of the wrist. The authors ascertained that the method is not recommended for females aged 17 and under although it has proven to be a reliable method for *male* football players in multiple studies [46–48]. Concluding, it is suggested to perform age estimation separate for gender to achieve the most reliable and accurate results for the specific target group.

The *socio-economic status* of the subjects of this work was similar to reduce possible variability. While this factor should be considered when evaluating a model, it is rather difficult in practice, to determine the exact background of an unknown subject, especially of asylum applicants. Regardless, EASO [57] and Schmeling et al. [173] state that low socio-economic status delays the skeletal maturation. Hence, applying a method trained on subjects of a higher socio-economic status, will most certainly underestimate the age of an individual with a lower status [57]. Hence, applying any method developed in the current work, would likely underestimate individuals with low socio-economic status. This outcome is in line with the principle

of "in dubio pro reo". Nevertheless, it is recommended to validate the method on other populations in the future to verify this hypothesis.

A further aspect to consider for the selection of a proper population is the *age range*. A large age range, e.g. 10 to 30 in [35, 52, 104], has the advantage that multiple age limits (14, 18, 21 [168]) can be evaluated with a certain margin of error. However, young individuals who are the beginning of adolescence or already grown-up, will exhibit limited growth plate ossification activity. This was observed for a small sample of the current study: all subjects from *Dataset A* younger than 15 years had completely open growth plates of the knee while all subjects above 21 had fully ossified growth plates (Fig. E.1). Moreover, considering that the most common legal ages in practice are between 14 and 21 years, an age range from 11 to 24 should be sufficient to determine the age of subjects within a margin of error around these limits of $\pm$ 3 years. The age range selected for the current work was rather "narrow" and did not enable the investigation of other legal ages such as 14 and 21 years. Nonetheless, the age range was suitable to evaluate the 18-year-limit with adequate margin of error.

The *age distribution* of the analyzed sample population is also critical for a reliable age estimation. A uniform distribution is favoured for an unbiased evaluation of each age group. This was a limitation of *Dataset A* and *Dataset B* of this work. The problem was addressed for *Method 2* and *Method 3* by augmenting the less represented age groups of the training set (Fig. 6.8). However, a fully-uniform distribution could not be attained. The results show higher errors for the less populated groups which suggests a probable bias (Fig. 7.16). This issue was part of the motivation for the acquisition of the retrospective dataset of sagittal MRIs (*Dataset C*). Merging the distributions from all three datasets (Fig. 3.2) created a nearly uniform distribution. The evaluation on the sagittal MRIs showed comparable results to the coronal data and even surpassed them in majority classification (sub-section 7.3). This supports the fact, that the rather "favourable" distribution in the coronal datasets was not the *sole* reason for the good results.

A total of 589 T1-weighted 3D MRIs of the knee were acquired in this work, but it is difficult to determine whether the amount is sufficient to capture the full variability in a population. Only minor other studies, such as [51, 52, 143, 207, 215], acquired 500 or more MRIs for age estimation. In general, it is challenging to access a high number of medical images due to law and policies that protect patient privacy [22]. Since age estimation is considered an "inexact science" [61], it is recommended to

acquire as much data as possible. Especially, considering the large amount of data required by deep learning approaches to effectively learn complex tasks. While it is generally recommended to expand the data as much as possible, the effort for age estimation solutions should be made in the acquisition of a *homogeneous* and *uniformly-distributed* population.

The final topic of this section is about the *image modality* selected for this work. The advantages of MRI is the capability of distinguishing bone from soft tissue. This allows a very detailed visual assessment of the growth plate maturation from an expert point of view. Moreover, the image modality is non-invasive which addresses one of the main issues, i.e. radiation, of methods currently used in practice. However, MRI has disadvantages and limitations concerning applicability in practice. For example, the cost of the modality due to the necessary equipment, spacious facilities, medical technical assistants, and examination duration is high in contrast to X-rays. With the steady advancement of technology it is likely that the cost for the equipment will decrease. Additionally, the MRI-scanners are expected to become smaller in the future which would decrease the costs for the facility. Another opportunity for cost reduction is in the acquisition times. In [100, 126, 140, 149, 167, 206, 218, 222, 225] the idea of image reconstruction using under-sampled data in k-space was proposed. This technique accelerates data acquisition in MRI examinations. Many of the studies make use of deep learning to learn spatio-temporal dependencies. This can not only reduce the cost but also minimize the stress to patients, especially young refugees with a traumatic background. In [140, 206] first promising results of this acceleration technique for age estimation were attained. Therefore, the suitability and applicability of the developed method in the current work will become more feasible in the future. The alternative would be to switch to other non-invasive image modalities such as ultrasound, but it has yet to be evaluated thoroughly for its usefulness in age estimation. Currently, EAS0 [57] considers ultrasound unsuitable for age estimation since it does not offer sufficient visualization of all growth plate OS.

Pre-Processing

N4ITK [211] proved to be an effective pre-processing method to correct intensity non-uniformities in 3D knee MRIs. However, the algorithm's implementation does

not account for noise in the image which is commonly present in MRIs. This shortcoming directly affected the bone segmentation step of this work (Fig. 7.6). Another important limitation of BFC is the time-consumption. The process required between two and ten minutes for the correction of previously downsampled 3D MRI. In contrast, the automated cropping, the segmentation, and the age estimation needed less than ten seconds each. There could be many possibilities to accelerate this process in the near future. First, BFC could be integrated in the acquisition procedure such that the radiologist or researcher is not affected by it. Second, deep learning methods could be used to distinguish non-uniformities and artefacts and use the gained knowledge to correct new images. Several studies have made significant progress for MRI data [45, 59, 188, 219]. Shaw et al. [188] proposed the idea to generate realistic artefacts, apply them to noise-free images, and ultimately train deep learning models on the artefact-afflicted images to increase the robustness of the models. Thus, the combination of accelerated image acquisition and noise correction through deep learning is a promising approach for the near future to mitigate cost and time-consumption and to improve the quality of medical images.

The *automated cropping* implemented in this work, successfully contributed to the standardization of the MRIs by extracting similar VOIs from images with different FOVs. Furthermore, the automated cropping reduced the data complexity for segmentation and age estimation by partially removing undesired anatomical structures and artefacts in the images. The reduction of the data had the further positive effect of increasing the fraction of the growth plates with respect to the whole image frame. The *automated cropping* was largely robust to different locations of the knee joint and distinct FOVs. An improvement for the future could be a parallelization of the algorithm to accelerate the process. Concluding, this approach is recommended to extract standardized VOIs in medical images and can surely be applied to other anatomical structures if the characteristic region for patch matching is redefined. If a deep learning solution is preferred, YOLO [152–154] is suggested for the cropping and detection task. Given the adequate hardware, it can potentially become faster than the patch matching of this work.

Segmentation

Due to the structural complexity and intensity variations of growth plates and bone it was not possible to satisfactorily extract these structures using established in-

tensity- or region-based approaches. The use of more advanced methods based on statistical intensity and shape models with subsequent atlas-based registration did not yield sufficient performance either. Especially for the detection of more complex bone shapes in image slices with a small amount of bone structures.

The aforementioned problem was solved using CNNs. The excellent DSC score of 98.5% proved that the designed network architecture (Fig. 5.8) was well-suited for the segmentation of knee bones in MRI. Table 8.1 compares the result from the current work with other studies. For the Femur, the DSC achieved in the current work was superior and for the Tibia, only Ambellan et al. [2] attained a higher similarity coefficient. No results on the Fibula, or on methods that segments all three knee bones, have been reported by others.

Table 8.1: Comparison of the performance of various segmentation models to other studies. Similarity metrics are presented in %.

Model type	DSC	IoU	Precision	Recall	Error
Femur	98.61	97.26	98.64	98.58	0.33
Femur [37]	94.00	-	94.60	93.90	-
Femur [64]	95.20	-	-	96.70	-
Femur [2]	98.50	-	-	-	-
Tibia	98.38	96.82	98.27	98.50	0.33
Tibia [39]	92.00	-	-	-	-
Tibia [30]	97.50	-	-	-	-
Tibia [64]	95.20	-	-	96.70	-
Tibia [2]	98.50	-	-	-	-
Fibula	96.19	92.66	96.81	95.59	0.07
Merged	98.50	97.04	98.53	98.47	0.72
Combined	97.26	94.67	99.58	95.05	1.29

In addition to achieving state-of-the art results in knee bone segmentation, the CNN architecture demonstrated robustness to noise applied in form of dilation and erosion to the training data (subsection 7.2). When trained anew on noisy data, the model predicted the bone structures in images of the test set with 28% less noise than it was trained on. This results suggests that the designed architecture is resilient to small errors that can occur during the generation of the gold-standard segmentation.

Furthermore, when applied to uncropped and larger coronal MRIs and on sagittal MRIs, the model made good predictions of bone structures in previously unseen

and unknown data (subsection 7.2). A shortcoming of merely *applying* the model to new data, was the insufficient capability of fully capturing the bone shafts. *Transfer learning* proved to be a successful technique to fine-tune a model on unknown data. In this work, the knowledge gained from the segmentation of coronal images was transferred to the task of segmenting sagittal MRIs. The LOOCV on a smaller sample of sagittal MRIs resulted in an average DSC of 97.42%, confirming the effectiveness of transfer learning given a small dataset. The performance can be improved further by integrating more data into the training or by amplifying the augmentation on available data.

To fully evaluate the segmentation on MR images of the knee, other CNN architectures were implemented. To exploit the 3D context of the knee MRIs, a CNN based on 3D convolutions was trained as well (subsection 5.7). The model achieves an average DSC of 98.28%, which is comparable to the 2D CNN. The downside of the 3D CNN was higher overfitting. Hence, more regularization in form of dropout layers had to be included in the network. Adding more data and conducting more augmentation could enhance the results of the 3D model.

A further architecture based on *dilated convolutions* was briefly tested. This type of convolution has recently gained popularity and has been used in a number of studies [134, 146, 221, 223]. The dilated CNN performed well and only had a marginally lower DSC of 98.34% compared to the merged model (appendix D). These results were not presented in 7.2 since no cross-validation was performed. Notwithstanding, the idea of using dilated convolution instead of down- and upsampling the images to capture features at different scales, is an interesting approach and could be considered for future analyses.

An observation from the network engineering phase was the relatively small impact on performance using different architectures. One reason could be due to *convolution*, which is the most important building block of CNNs. It is translation invariant and results in a more efficient image analysis [23]. Hence, less samples are generally needed to learn representations that fully capture the variability in the data. Another reason could have been the image data itself. Cicek et al. [25] state that a network trained on medical images has a high generalization power even when trained on few samples. This can be attributed to the fact that neighbouring slices show very similar information. Consequently, it could be assumed that the acquisition of enough data can cause well-designed but different CNNs, to attain a relatively good and comparable performance. Lastly, pre-processing steps can eclipse any changes

performed in the network architecture [112]. This assumption was confirmed during the development phase of the current work. In [148] the merged model was trained on a smaller sample, the extraction of standardized VOIs relied on intensity distribution instead of patch-matching, and the augmentation did not include the generation of different FOVs. The DSC score, in addition to the other segmentation metrics, was 98% in comparison to 98.5% for the latest results. The network architecture and it's hyperparameters remained unchanged.

There are also some limitations of the implemented approach. First, the trained merged model had difficulties handling strong MR artefacts. This can be mitigated by adding, or synthetically generating, noise to the training data as previously mentioned. Second, augmentation in the current work only performed rigid transformations such as translation and rotation. Elastic transformation as a type of augmentation would offer a further possibility to increase the training data and robustness of a CNN [189]. Third, augmentation was conducted prior to training. "Online" augmentation could be used instead. It could potentially result in a more robust model since the training data is randomly transformed after each training epoch. Therefore, more variation in the training data is generated. Finally, segmentation quality could be improved further by introducing anatomical prior knowledge into CNNs [2, 141].

Age Estimation

The proposed solution for age estimation addresses many of the shortcomings of actual methods. It is based on a non-invasive image modality, it is computer-based an fully automated, and is based on an actual and homogeneous population. In contrast, current methods in practice still rely on the visual assessment of the growth plate ossification degree by a radiologist. This is subjective and prone to error and requires the predefined OS to provide enough criteria to estimate the age of a young individual and to discriminate between adults and minors.

To verify, if the definition of ossification offers sufficient foundation for majority classification, a similar analysis to the established methods was performed for *Dataset A* [3]. The earliest age of full ossification of the femoral growth plate that was observed, was at the age of 16.3 years (appendix E). The sole practical observation from the analysis was that all subjects with an ossification stage I or an SKJ < 5 were minors

(Fig. E.1). In contrast, in [104, 162] a complete ossification of the distal femoral epiphysis did not occur before the completion of the 18th year of life. The results are conflicting, suggesting that the approach by visual inspection is not reliable for classification. Therefore, a solution similar to the proposed method of the current work is advised.

Regarding the proposed method for age estimation, the initial idea to train a CNN on the original MRIs was unsuccessful (Fig. 7.14). The training of the network was unstable and the final age regression led to an average MAE of 0.97 ± 0.84 years (Tables E.2 and E.5). Similar results were computed for the CNN trained on unmasked *sagittal* MRIs (Table E.6). These predictions translate to an average maximal deviation to the actual chronological age of 2.63 years or more for 95% of the samples and is worse the ± 2 years reported in literature for methods used in practice. For the exact same fold, the CNN based on *masked images* achieved an average MAE of 0.82 ± 0.64 years. This is an important improvement and represents a maximal deviation of 2.1 years, which is in line with reported values in literature. The gain through segmentation supports its use as a preprocessing step to extract age-relevant structures from the images and to transfer knowledge to the model trained for age estimation.

The CNN results based on masked images were enhanced further by using ML algorithms in a second step. These algorithms effectively trained on the age predictions made by the CNN per image slice and delivered better results (Tables 7.5, 7.6, 7.7). Moreover, the combination of CNNs and ML-based regressors was superior in comparison to the use of 3D CNNs to regress the age of an individual (Tables E.5 and E.6). Nevertheless, a 3D CNN has the potential to gather more contextual information from multiple slices and should be analyzed in more detail in the future.

The incorporation of additional features such as AM and OS into the ML algorithms, only marginally boosted the performance on age regression (Table 7.5). Furthermore, the CNN that was generated to handle those features as additional inputs (section 6.3) delivered worse results (Table 7.7) in comparison to integration of the features in the ML algorithms. However, more time could be invested in the future to adapt and improve a *multi-input* and *mixed-data* model for age estimation. Such a network type, enables the possibility for an *end-to-end* training.

ML algorithms trained solely on the numeric data acquired, i.e. the AM and OS, had higher predictions errors in comparison to models based on *Method 2* or *Method 3*.

Yet, the combination of AM and SKJ showed potential and attained a MAE of 0.77 ± 0.60 years. A feasible option for future analysis, could be the incorporation of further data, such as psychological assessments and sexual maturation, to train ML algorithms on age regression. These algorithms have the advantage that they are simple and fast to train.

The final comparison of age regression methods is between coronal and sagittal MRIs. The performance of *Method 2* is better on coronal than on sagittal data. This could be related to a higher possibility of variance and outliers of the sagittal dataset due to its larger size. Additionally, the age range was broader (by 1 year) and the age distribution was more uniform. Both facts could have induced further variance. Similar conditions should exist to make a final comparison.

In regards to the classification of the 18-year-limit, comparable results were achieved using coronal or sagittal MRIs (Tables 7.9 and 7.10). For the latter, the accuracy, sensitivity, specificity, and a AUC were slightly higher. An intriguing observation from the sagittal results is that the classification performance is better in comparison to the coronal case, even though the age predictions from the CNN were worse in the regression task. A possible explanation could be the statistical impact due to a larger number of samples. An important consideration for the future is to focus on improving the sensitivity. EASO [57] encourages to follow the principle of "in dubio pro reo" and hence falsely classified minors should be mitigated.

Method 1 had a lower performance on majority classification in comparison to *Method 2* and *Method 3*, with all metrics around 80% (Table 7.8). Surprisingly, the combination of AM and OC into a classifier did not improve the results as it was the case for age regression. It was not possible to determine the cause for this behaviour.

Finally, the incorporation of AM and OS into the classifiers of *Method 2* for coronal knee MRIs, did not generate a substantial gain. It improved the specificity of the models at the cost of sensitivity (Table 7.9). Future analyses should evaluate whether all the data is necessary to classify minors and adults. A *principal component analysis* (PCA) or similar techniques could bring insight to this matter.

Next, the proposed method is compared to the works by Stern et al. [195–199]. This research group has developed a method for age estimation based on machine learning and deep learning using MRI modality as well. One of the main differences to their works is the investigated anatomical site. In [195, 198, 199] they used 3D

MR *hand* images and in [196, 197] they expanded to a multi-factorial data, including MRI volumes of the *hand*, *clavicle*, and *wisdom teeth*. Table 8.2 enlists the results on regression.

Table 8.2: Comparison of age regression performance between the current work and other studies

Study	N	Gender	Age Range [y]	Anatomical Site	MAE ± SD
M2-COR*	185	Male	14-21	Knee	0.67 ± 0.49
M2-SAG*	404	Male	13-21	Knee	0.79 ± 0.57
[195]	56	Male	13-19	Hand	0.85 ± 0.58
[199]	132	Male	13-20	Hand	0.82 ± 0.56
[196]	103	Male	13-24	Hand, Teeth, Clavicles	1.14 ± 0.96
[197]	322	Male	13-25	Hand, Teeth, Clavicles	1.01 ± 0.74
[198]	328	Male	13-25	Hand	0.82 ± 0.65

∗ Models based on *Method 2* (M2) using coronal (COR) or sagittal (SAG) MRIs

The similarities to the studies by Stern et al. are the gender and age range. These are some of the most critical factors when comparing studies on age estimation and is often a considerable problem. The number of datasets varies but [197, 198] are comparable to the sagittal dataset of the current work ($n = 404$) and [199] with the coronal dataset ($n = 185$). The major difference, as mentioned above, is the investigated anatomical site. While it makes an unbiased comparison more difficult, it also offers the opportunity to determine the importance and potential of the anatomical site for age estimation. Both M2-COR and M2-SAG surpass the results from the studies by Stern et al.. However, the comparison of the methods should be taken with caution. In [195] and [199] the authors solely used a random forest regressor for age regression, while in the other two works they developed a deep convolutional neural network for multi-factorial age estimation. Additionally, the CNN architecture is only partially comparable to the proposed one of the current work. They focussed on 3D convolutions, fused architectures for the three anatomical sites, and in [196], pre-trained the model on the radiological assessment of the maturation of the growth plates. Similarly to the current work, Stern et al. also observed an improvement of age regression through pre-training.

Related to the majority classification, the comparative results can be found in Table 8.3. All metrics of Stern et. al. are noticeably similar and the models perform

well on the classification task. The difference to the current study is the balance between minors and adults. Stern et. al. have a greater amount of adults in the population which could be an explanation of the high specificities and rather low sensitivity in [197]. Notwithstanding, the AUC in the last-mentioned study is remarkably high with 98%.

Table 8.3: Comparison of majority classification performance between the current work and other studies.

Study	Minors (%)	Accuracy	Sensitivity	Specificity	AUC
M2-COR*	49.25	89.71	88.18	92.31	91.99
M2-SAG*	52.26	90.93	88.64	94.19	94.38
[196]	42.72	91.30	88.60	93.20	-
[197]	41.62	90.68	82.10	96.80	98.00

∗ Models based on *Method 2* (M2) using coronal (COR) or sagittal (SAG) MRIs

9 Conclusions

The current work presents a new computer-based approach for the automated age estimation of young individuals using 3D knee MRIs. The approach consists of three main steps: the pre-processing of the MRIs, the subsequent extraction of age-relevant structures (bones), and ultimately, the estimation of the chronological age based on the extracted bones.

As part of the main contributions of this work, is the *automated cropping* as a pre-processing step to extract standardized VOIs in knee MRIs, irrespective of the FOV and anatomical position of the knee joint. It is a robust technique that could easily be adapted for similar approaches by selecting a task-related characteristic image region.

The CNN-based segmentation of this work achieves state-of-the-art results in the detection of bones and has a good performance when applied to unseen MRIs of the knee in different sizes and orientations. An approach similar to the retraining performed for sagittal knee MRIs via *transfer learning* is suggested to solve similar problems with a manageable effort.

Finally, the new automated age estimation method proves its capability for both the regression and classification tasks. The recommended approach for future applications is *Method 2*, which combines a CNN and an ML algorithm. For regression, the extremely randomized trees regressor and for classification, the random forest classifier shows the greatest potential.

The proposed age estimation method of this work will improve further and become more reliable in the future when trained and validated on large and diverse datasets.

A Hardware and Software

The essential hardware of the workstation used for this work were the CPU, the random-access memory (RAM), and especially important for deep learning, the graphics processing unit (GPU). The details of these components can be found in Tab. A.1.

Table A.1: Essential hardware available for this work

Hardware type	Quantity	Model name
CPU	1	Intel® Xeon® CPU E5-1650 v4 (3.60 GHz)
GPU	4	GeForce® GTX 1080 Ti
RAM	4	Kingston® 16GB DDR4 2133 MHz RAM

The programming environment was set up in the Linux operating system "Ubuntu 16.04.6 LTS". The software and scripts were written in the programming languages (PL) Python (version 3.5.2) and C++ (g++ 4.8.4 compiler). The most important Python and C++ libraries and frameworks for this work are enlisted in Table A.2.

Table A.2: Most important Python and C++ libraries and frameworks

Name	Version	PL	Description
ITK	4.9.0	C++	Open-source software for image analysis
Keras	2.2.4	Python	Open-source neural-network library
matplotlib	3.0.2	Python	Plotting library
numpy	1.15.4	Python	Package for scientific computing
pandas	0.23.4	Python	Open-source data analysis library
scikit-image	0.14.1	Python	Open-source image processing library
scikit-learn	0.20.0	Python	Machine learning library
scipy	1.1.0	Python	Open-source scientific computing library
seaborn	0.9.0	Python	Data visualization library
SimpleITK	1.1.0	Python	Open-source interface to ITK
tensorflow	1.5.0rc0	Python	Open-source machine learning library
VTK	7.0.0	C++	Open-source software for image analysis

To fully exploit the potential of deep learning approaches, the workstation was set up with modern GPUs (Table A.1). CUDA[1], a parallel computing architecture from NVIDIA, and cuDNN[2], a GPU-acclerated library for deep neural networks, were installed to access the GPU capabilites of Keras[3] and TensorFlow[4]. CUDA version 8.0.61 and cuDNN version 6.0.21 were used.

The sources of this work will be made available in a Bitbucket[5] repository in the near future. To request access to the repository, contact can be made through the following e-mail: `markusalexander.adm@gmail.com`.

[1] `https://www.nvidia.com/object/io_69526.html`
[2] `https://developer.nvidia.com/cudnn`
[3] `https://keras.io/`
[4] `https://www.tensorflow.org/`
[5] `https://bitbucket.org/product`

B Overview of MR Artefacts

MR imaging has evolved significantly over the last decades, improving the quality of diagnosis and treatment with the information extracted from its data. MRI can generate data showing various types of tissues without the use of radiation. However, in comparison to other imaging modalities, such as CT and ultrasound, it is more susceptible to various kinds of artefacts, especially patient motion [224]. The problem is that any type of movement, voluntary or not, is faster than the time required to collect enough data to create an image [224]. The issue is present in most available MRI sequences. Motion and other artefacts can have a negative influence on the quality of diagnosis [54, 107] as well as on the performance of image processing methods, e.g. segmentation, classification, registration, and texture analysis [8, 9, 84, 94, 200]. Commonly, these artefacts are divided into three groups [53, 54]:

- patient-related artefacts
- signal-processing artefacts
- hardware/machine-related artefacts

Patient-related artefacts. Motion is the most common artefact overall. It can be periodic or bulk (i.e. rigid). Periodic motion, e.g. due to blood flow, cardiac and respiratory motion, creates discrete *ghost artefacts* along the phase-encode direction. "Ghosting" is a partial or complete copy of a structure or object. In contrast, bulk motion results in diffuse noise spread broadly along the phase-encode direction. Further patient-related artefacts can occur when metal objects are present in the patient's body. These can cause image distortion, high signal, and even signal loss. [53, 54, 107, 224]

In the current work, no subjects had any metal objects inside their bodies. Periodic motion was not observed and is generally not an issue in knee MRIs. Bulk motion of the subjects was limited due to the knee coil used, but could not be completely avoided (Fig. B.1).

Figure B.1: Motion artefacts in knee MRIs

Signal-processing-related artefacts. These can further be separated into chemical shift, partial volume, wrap around, Gibbs phenomenon, and more, artefacts [53, 54].

Chemical shift artefacts mainly occur at the interfaces between water and fat due to the difference in resonant frequency of the protons of these two tissues. This will cause a shift in the mapping of fat and water pixels when the image is created. The chemical shift is present in form of dark and bright bands, often left and right of the tissue boundaries. The effect of this artefact is intensified with increasing magnetic field strength of MR machines. In MR images of the knee, as the ones from the current study, the chemical shift can appear at the edges of bones, causing difference in cartilage thickness. [53, 54, 107]

Partial volume effects occur when the signal from different tissues in a voxel are averaged out and will thus cause a loss of spatial resolution. These effects can be mitigated by choosing smaller pixels/voxels and/or a smaller slice thickness. [53, 54]

Another artefact related to signal-processing is the *(phase) wrap-around* or *aliasing artefact*. It occurs when the size of an anatomical structure is larger than the chosen FOV. This will fold the parts outside the FOV to the opposite side of the image. One possible solution is to increase the FOV and another one to perform phase oversampling, which is often already available in the MR-scanner software. [53, 54, 107].

Several of the knee MRIs of this work showed overlapping structures, but almost exclusively the ones acquired with 1.5T MR-scanners (Fig. B.2).

The last of the most common signal-processing artefacts is the *Gibbs phenomenon* or truncation/ringing artefact. It can appear in form of alternating and evenly spaced dark and bright bands close to sharp high-contrast boundaries. [53, 54, 107]

This effect was seen in some of the knee images, but generally just in individual slices and not on the entire dataset (Fig. B.3).

Figure B.2: Wrap-around artefacts observed in knee MRIs

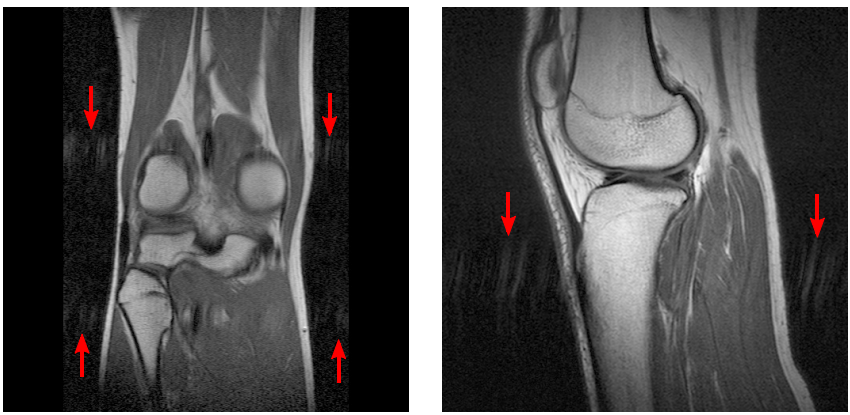

Figure B.3: Ringing artefacts

Hardware/machine-related artefacts. This third and final group of artefacts is related to the MR machine and its components. Inhomogeneities of the external, i.e. static, magnetic field B_0 and the gradient fields B_1 cause intensity inhomogeneities in the MR images [8, 9, 54, 84, 94]. To understand how these field inhomogeneities affect the outcome of the images, a brief background will follow next.

The static magnetic field causes hydrogen protons in the body to align and precess around its field direction. The gradient coils on the other hand, can be switched on to temporarily create perpendicular and oscillating fields, which cause a frequency variation of the protons along the direction of the gradient. When switched off again, the protons gradually align back towards the static field. The excitation of the protons will emit radio frequencies which can be measured. The use of three types of gradient fields (x, y, and z) allows the spatial encoding of the MR signal and finally the generation of the images. Thus, any inhomogeneity in either the static or gradient fields, will not evenly excite the protons at the desired location. The resulting images will exhibit spatial and/or intensity distortions (Fig. B.4). [53, 54]

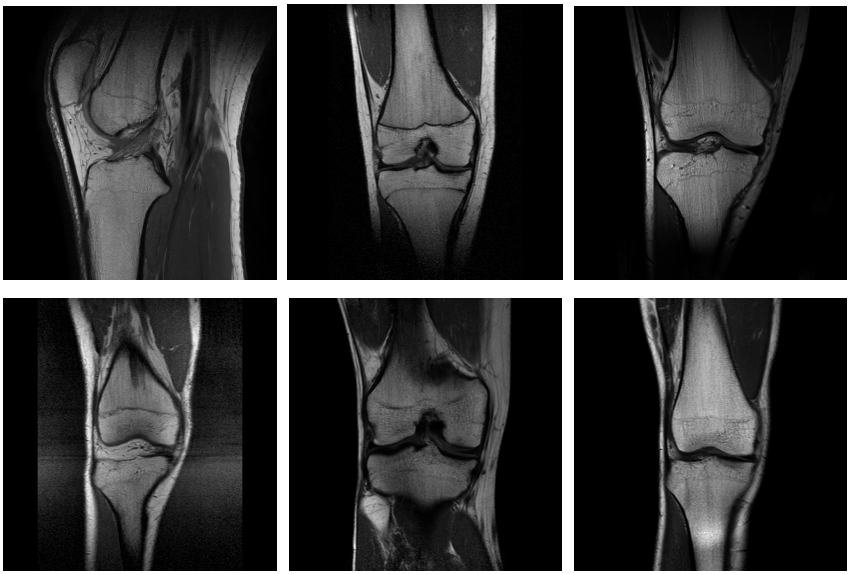

Figure B.4: Intensity distortions

Many of the artefacts described above can only be mitigated or avoided at the time of the image acquisition. The ones related to magnetic field inhomogeneities can be corrected using an approach called *Bias Field Correction* which was introduced as a pre-processing step for the data of this work (section 4.2). A popular algorithm for BFC is N4ITK [211] and was used in this work.

C Augmentation

The augmentation of the knee images of this work had to be refined in comparison to the initial approach describe in [148]. When translation and rotation is applied to the already cropped images from section 4.3, parts of the structures exit and new pixels enter the image frame (Fig. C.1 – bottom row). The new pixels are not known in the cropped MRIs and therefore filled with zeros which causes the loss of information. However, in the full size images this information is available in most cases. Therefore, augmentation was performed *prior* to the actual cropping to recover "lost" anatomical structures (Fig. C.1 – top row).

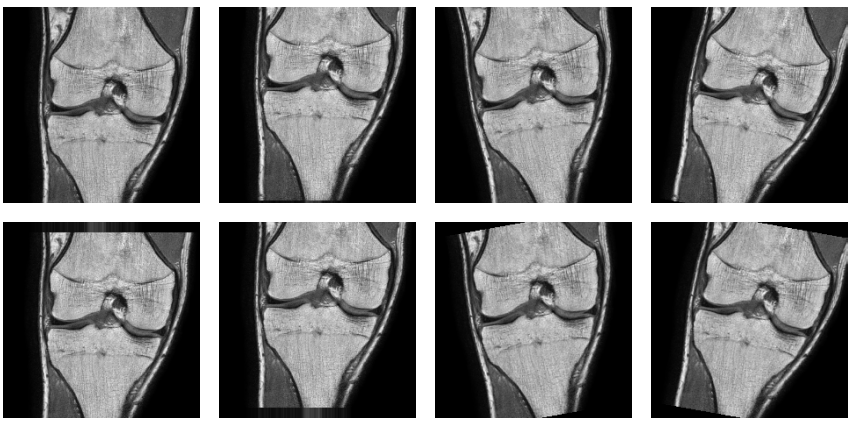

Figure C.1: Comparison of augmenting images before cropping (top row) and after cropping (bottom row)

D Further Results on Segmentation

Segmentation Quality

Image segmentation quality in the sense of traditional computer vision is often measured for each 2D image separately using DSC score. For this purpose, the DSC was computed separately for each of the 435 coronal image slices of a test set using the predictions of the corresponding merged model (Fig. D.1).

The average DSC over all 435 images of the test set was 0.965 ± 0.103. These results are slightly more pessimistic than the ones in Table 7.3 since one false prediction has more impact on a single 2D image (224×224) in comparison to the whole test set ($435 \times 224 \times 224$). From the distribution in the figure a few "outliers" results can be seen with a DSC between 0.0 and 0.05 which negatively influence the

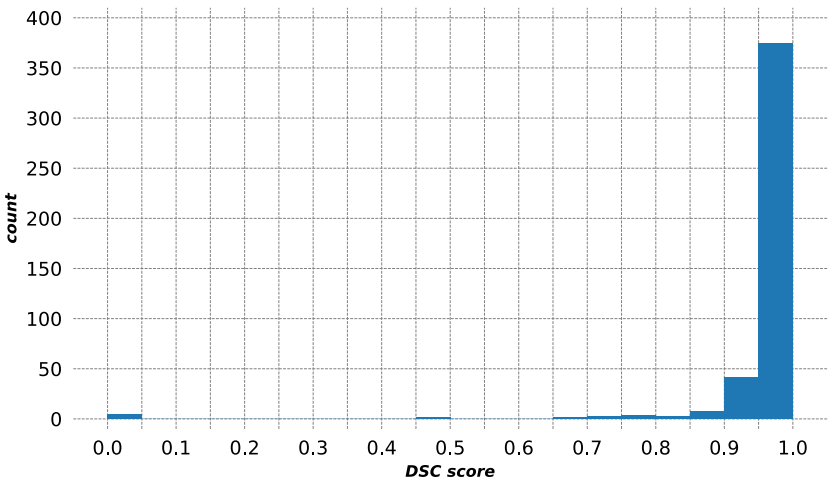

Figure D.1: Histogram showing the DSC score distribution for one trained merged model on coronal MRI data. The average DSC over the 435 samples of the test set was 0.97 ± 0.10. The median value is 0.987, showing the influence of a few bad predictions on the mean.

average score. The median DSC of the aforementioned distribution was 0.987 and is more comparable to the results acquired for the overall segmentation performance (subsection 7.2). This confirms the influence of a few "bad" segmentations on the average score.

Dilated Convolutions

Dilated convolutions have been used in a variety of applications instead of scaling operations [134, 146, 221, 223]. A network similar to the one from Moeskops et al. [134] was implemented for the available knee MRIs.

The modifications included the use of Dropout (0.2) after convolution layers (except after the first one), the exclusion of Batch-Normalization, the selection of sigmoid activation for the final layer, and the reduction of the number of convolutional layers to nine. The number of kernels per convolutional layer was set to 32 as well and kept constant, the dilation scheme was identical, the activation function was ELU. For training, the optimizer was Adam with a learning rate of 0.001, a F_1-loss function, a batch size of 48, and 50 training epochs.

The dilated CNN was trained five times for a single fold. The average performance resulted in a DSC of 98.3%, an IoU of 96.7%, a Precision of 98.2%, a Recall of 98.5%, and a total prediction error of 0.8%. The loss variations between training rounds were negligible and no overfitting was observed.

E Further Results on Age Estimation

Assessment of Growth Plate Ossification

Age estimation based on the "traditional" visual assessment of the growth plate ossification degree by experts was investigated in this work as well. The three-stage system by Jopp et al. [92] was applied on all three growth plates of the knee for the subjects of *Dataset A* and *Dataset B* and is described in section 3.4.

In this section the focus is on *Dataset A* which is a longitudinal study and enabled the analysis of the ossification progression over time. The complete and detailed analysis was published in the International Journal of Legal Medicine [3].

The relationship between the chronological age of subjects from *Dataset A* and the ossification classes of the three knee bones and the SKJ based on coronal MRIs was analyzed (Fig. E.1). The boxplots in the figure show the median age as a line (orange), the box extending from the lower to upper quartile, the whiskers from ± 1.5 IQR, and the outliers as circles. The left sub-figure shows that all subjects with class I, in any bone, were younger than 18 years. However, classes II and III extended below the 18-year-limit (red line) which means that this method was not suitable to correctly classify all underage subjects. Moreover, none of the ossification classes was unique for adults. Similar statements can be made for the right sub-figure. An SKJ < 6 indicated that the subject was minor. In contrast to OS, SKJ offered additional information concerning all growth plates: score 4 and 5 indicated that a subject could still be underage even if one or two of the three growth plates of the knee was partially closed. Nonetheless, all other scores also included minors and again no SKJ value was unique for adults.

Given the longitudinal nature of *Dataset A*, the temporal aspect of the ossification could be analyzed. For this investigation the change in SKJ was observed and accumulated over a 2-year period (from BL to FU2) for 29 subjects of *Dataset A* with exactly three MRI examinations (Fig. E.2). The average values for each age at the first time-point (BL) are drawn as dots and the red vertical lines represent ± 1σ. The ossification activity increased steadily between 14 and 16 years with a peak

of almost four SKJ steps on average at the latter age. Subjects older than 16 at the start of the study showed less change in SKJ which means that their ossification status was already more advanced at that time-point.

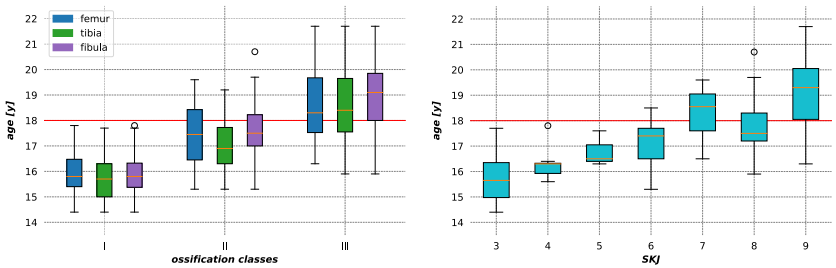

Figure E.1: Age vs. ossification classes (left) and vs. SKJ (right) acquired by visual assessment of the subjects from *Dataset A*.

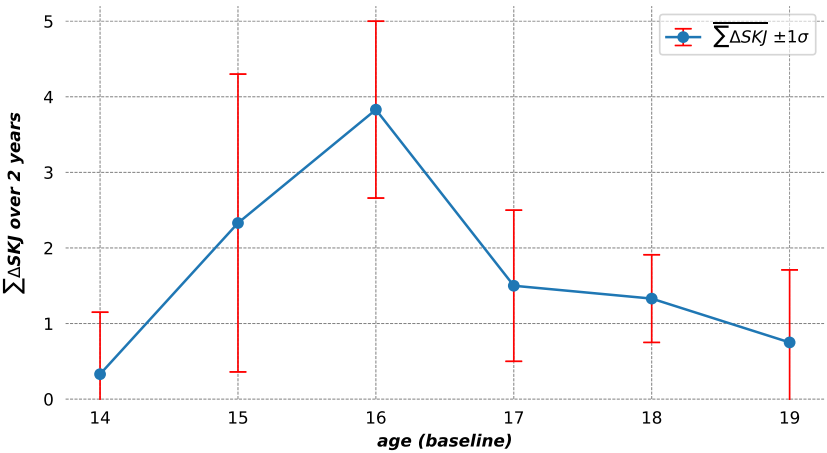

Figure E.2: Change in SKJ accumulated over a 2-year period for 29 subjects from *Dataset A*. The average values with 1 standard deviation are plotted for subject ages at the first time-point, i.e. baseline. A peak can be observed around 16 years indicating the most active time-point of growth plate ossification in the knee.

Other observations based on the longitudinal dataset was the distinct ossification patterns for each growth plates of the knee. The proximal tibial epiphysis matured earlier compared to the other two epiphyses in many cases. Other studies gathered similar insight but based on single time points [35, 52]. The current work can

confirm these results based on intra-individual development. Furthermore, it was noticed that the ossification pace of the tibial epiphysis was slower and but more continuous in comparison to the femoral one. There were many cases when the femoral epiphysis had a jump of two OS within one year. Contrarily, the Fibula showed signs of a steady growth plate maturation and finished the process after the other two knee bones on average.

Distribution of Model Prediction Errors

In another analysis on age regression, the distribution of the model prediction errors were observed (Fig. E.3). The figure shows the error distributions of the best performing model from *Method 2* before (left) and after (right) applying the final regression using the ML algorithms. The age prediction errors of the model reduce when using the ML models (right): the errors improve from the range of -4 to 5 years to the range of -3 to 3 years and the frequencies around zero increase. A further interesting observation from the plots is that there appear to be slightly more errors on the negative side. This means that the age of subjects was more often underthan overestimated using this approach and would favour the principle of "in dubio pro reo".

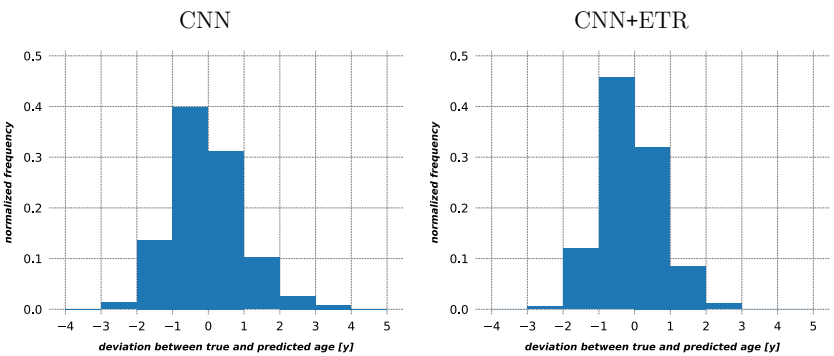

Figure E.3: Distribution of the age deviation between predicted and actual chronological age from a CNN of *Method 2* without (left) and with ML models to regress the final age (right). Averaged for the best model of each of the five folds.

Occlusion Method

A further analysis on the age estimation task, more specifically of *Method 2*, was to observe which parts of the knee images were *possibly* more relevant for age estimation. The "occlusion method" described by Zeiler and Fergus [226] was used for this purpose. It is generally used for classification and hence, had to be adapted to the age regression task. The goal of the occlusion method is, as the name suggests, to occlude part of the image and subsequently observe the effect it has on the prediction of the trained model.

In a sliding-window fashion, the absolute age difference between the prediction of the CNN with and without the occlusion is computed at each point. The output of the method was exemplary generated for three images with different ossification degree (Fig. E.4). The left column shows the image slice with the occlusion patch, the middle one the heatmap of the absolute change in prediction, and the last one the heatmap as overlay on the image slice. The patch was selected rather large to verify if the CNN focuses on bone or background. The heatmaps appear to indicate that the central region of the image is more important (red) for the task than the outer areas (purple). The areas in red mostly cover bone structure close to the knee joint with a higher focus on the proximal end of the tibia. The results from the occlusion suggest that bone structures close to the knee gap are more important for the age estimation task based on a CNN architecture as described in section 6.2. However, the results should be interpreted with caution. A simple change in the size of the occlusion can shift the important regions and constrain them locally making the heatmaps harder to interpret.

Regression and Classification

Following, the complete collection of results for models of *Method 1*, *Method 2*, and *Method 3* on age regression (Figures E.1, E.2, E.3, E.2, E.4, E.5, E.6) and majority classification (Figures E.7, E.8, E.9, E.10).

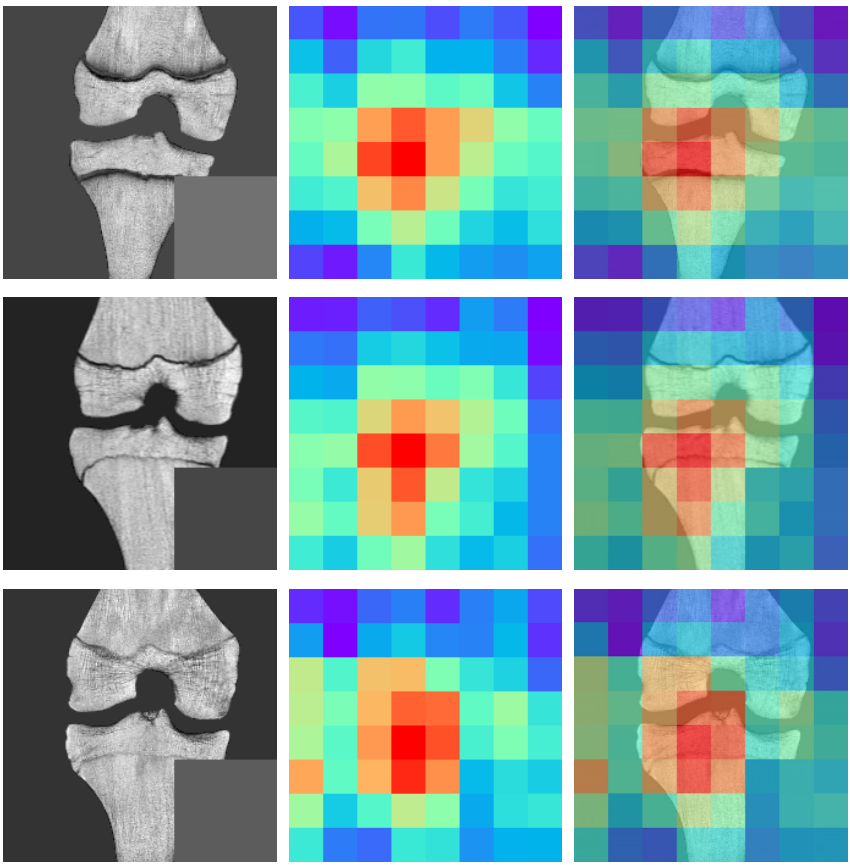

Figure E.4: "Occlusion method" [226] used to visualize the regions in the knee MRIs which have a higher (red) or lower (purple) impact on the prediction of a CNN for age estimation.

Table E.1: Performance of multiple M1 models on *age regression* in an "extended" 5-fold cross-validation using anthropometric measurements (AM) and ossification stages (OS). Metrics were averaged over multiple training rounds and five folds.

Model	MAE ± SD	Max AE	% ≤ $\|1.0y\|$	% ≤ $\|2.0y\|$
stat*	1.23 ± 0.90	3.93	48.53	79.20
M1-AM-GBR$_{def}$	1.00 ± 0.70	2.63	56.57	90.93
M1-AM-SVR$_{opt}$	1.00 ± 0.69	2.57	54.29	92.00
M1-OS-GBR$_{def}$	0.91 ± 0.68	3.01	60.00	92.00
M1-OS-GBR$_{opt}$	0.90 ± 0.68	3.00	65.71	92.57
M1-SKJ-ETR$_{def}$	0.91 ± 0.68	3.01	64.00	92.00
M1-SKJ-GBR$_{opt}$	0.90 ± 0.68	2.98	65.71	92.57
M1-AM-SKJ-ETR$_{def}$	0.90 ± 0.71	3.13	64.34	92.51
M1-AM-SKJ-RFR$_{def}$	0.87 ± 0.66	2.64	64.69	93.26
M1-AM-SKJ-GBR$_{def}$	0.84 ± 0.63	2.62	66.11	94.86
M1-AM-SKJ-SVR$_{def}$	0.85 ± 0.64	2.68	69.14	93.14
M1-AM-SKJ-LR$_{def}$	0.87 ± 0.68	2.88	66.29	92.00
M1-AM-SKJ-GBR$_{opt}$	0.80 ± 0.61	2.49	68.51	94.51
M1-AM-SKJ-SVR$_{opt}$	0.77 ± 0.60	2.58	72.00	95.43

*: static prediction of the mean age of the training set
def/*opt*: machine learning algorithms with default and optimized parameters
OS/*SKJ*: ossification stages via three-stage system by Jopp et al. [92] or via *SKJ*

Table E.2: Performance of multiple M2 models on *age regression* in an "extended" 5-fold cross-validation using *coronal* MRIs. Metrics were averaged over multiple training rounds and five folds.

Model	MAE ± SD	Max AE	% ≤ $\|1.0y\|$	% ≤ $\|2.0y\|$
stat*	1.63 ± 0.99	3.59	28.66	59.70
M2-NPT-NOSEGM**	0.97 ± 0.84	4.76	62.22	88.47
M2-NPT$_{all}$	0.85 ± 0.67	3.71	67.77	93.15
M2$_{all}$	0.81 ± 0.65	3.55	69.57	94.00
M2-NPT$_{best}$	0.81 ± 0.64	3.62	69.24	94.43
M2$_{best}$	0.79 ± 0.62	3.49	71.14	95.05
M2-ETR$_{all-def}$	0.73 ± 0.56	2.50	74.42	96.46
M2-LR$_{all-def}$	0.76 ± 0.58	2.59	74.80	95.54
M2-SVR$_{all-def}$	0.73 ± 0.56	2.52	75.03	96.51
M2-ETR$_{best-def}$	0.67 ± 0.49	2.10	78.86	98.86
M2-LR$_{best-def}$	0.71 ± 0.52	2.42	78.86	96.00
M2-SVR$_{best-def}$	0.70 ± 0.54	2.40	77.71	98.29
M2-ETR$_{all-opt}$	0.71 ± 0.55	2.46	75.17	96.78
M2-SVR$_{all-opt}$	0.71 ± 0.56	2.56	76.34	96.69
M2-ETR$_{best-opt}$	0.69 ± 0.51	2.30	77.71	97.14
M2-SVR$_{best-opt}$	0.69 ± 0.53	2.47	78.29	98.29
M2-ETR-FEATS$_{all-def}$	0.73 ± 0.56	2.47	74.44	96.66
M2-LR-FEATS$_{all-def}$	0.75 ± 0.57	2.57	75.94	96.00
M2-SVR-FEATS$_{all-def}$	0.72 ± 0.56	2.52	75.20	96.74
M2-ETR-FEATS$_{best-def}$	0.69 ± 0.47	2.15	77.71	98.29
M2-LR-FEATS$_{best-def}$	0.70 ± 0.52	2.44	80.00	98.29
M2-SVR-FEATS$_{best-def}$	0.69 ± 0.53	2.45	78.86	97.71
M2-ETR-FEATS$_{all-opt}$	0.71 ± 0.57	2.56	76.00	96.47
M2-SVR-FEATS$_{all-opt}$	0.73 ± 0.55	2.39	73.54	97.60
M2-ETR-FEATS$_{best-opt}$	0.70 ± 0.53	2.34	75.43	97.71
M2-SVR-FEATS$_{best-opt}$	0.68 ± 0.53	2.45	77.71	98.29

∗: static prediction of the mean age of the training set
∗∗: only performed 10 times for a single fold
npt: not-pretrained with the weights from segmentation task
def/*opt*: machine learning algorithms with default and optimized parameters
all/*best*: all or best training rounds of each fold included
feats: inclusion of anthropometric measurements and SKJ into MLMs

Table E.3: Performance of multiple M2 models on *age regression* in an "extended" 5-fold cross-validation using *sagittal* MRIs. Metrics were averaged over multiple training rounds and five folds.

Model	MAE ± SD	Max AE	% ≤ $\lvert 1.0y \rvert$	% ≤ $\lvert 2.0y \rvert$
stat*	1.93 ± 1.20	4.74	27.69	54.98
M2$_{all}$	0.92 ± 0.73	4.31	62.83	90.91
M2$_{best}$	0.89 ± 0.70	4.22	63.94	92.44
M2-ETR$_{def-all}$	0.82 ± 0.64	3.02	68.38	93.88
M2-LR$_{def-all}$	0.82 ± 0.63	2.95	67.79	94.16
M2-SVR$_{def-all}$	0.82 ± 0.63	2.87	67.92	94.00
M2-ETR$_{def-best}$	0.79 ± 0.57	2.63	70.67	95.73
M2-LR$_{def-best}$	0.81 ± 0.59	2.84	68.27	95.73
M2-SVR$_{def-best}$	0.80 ± 0.59	2.78	67.73	94.93
M2-ETR$_{opt-all}$	0.81 ± 0.62	2.92	69.26	94.63
M2-SVR$_{opt-all}$	0.81 ± 0.62	2.86	68.40	94.85
M2-ETR$_{opt-best}$	0.79 ± 0.58	2.82	70.40	95.73
M2-SVR$_{opt-best}$	0.79 ± 0.59	2.76	68.53	95.73

∗: static prediction of the mean age of the training set
all/*best*: all or best training rounds of each fold included
def/*opt*: machine learning algorithms with default and optimized parameters

Table E.4: Performance of multiple M3 models on *age regression* in an "extended" 5-fold cross-validation using *coronal* MRIs, AM, and OS. Metrics were averaged over multiple training rounds and five folds.

Model	MAE ± SD	Max AE	$\% \leq \lvert 1.0y \rvert$	$\% \leq \lvert 2.0y \rvert$
stat*	1.63 ± 0.99	3.59	28.66	59.70
M3$_{all}$	0.92 ± 0.70	3.47	62.61	91.49
M3$_{best}$	0.85 ± 0.64	3.29	65.62	94.33
M3-ETR$_{def-all}$	0.86 ± 0.66	2.80	66.66	92.46
M3-LR$_{def-all}$	0.88 ± 0.65	2.72	64.00	92.51
M3-SVR$_{def-all}$	0.86 ± 0.65	2.70	66.11	92.29
M3-ETR$_{def-best}$	0.71 ± 0.54	2.20	75.43	95.43
M3-LR$_{def-best}$	0.77 ± 0.56	2.42	70.86	96.57
M3-SVR$_{def-best}$	0.76 ± 0.58	2.39	73.14	95.43
M3-ETR$_{opt-all}$	0.84 ± 0.65	2.74	67.52	92.78
M3-SVR$_{opt-all}$	0.88 ± 0.66	2.77	65.03	92.74
M3-ETR$_{opt-best}$	0.75 ± 0.55	2.28	74.29	95.43
M3-SVR$_{opt-best}$	0.77 ± 0.58	2.59	72.57	96.00

$*$: static prediction of the mean age of the training set
def/*opt*: machine learning algorithms with default and optimized parameters
all/*best*: all or best training rounds of each fold included

Table E.5: Performance of other age regression models using coronal MRIs. Metrics were averaged over multiple training rounds of a single fold.

Model	MAE ± SD	Max AE	% ≤ $\lvert 1.0y \rvert$	% ≤ $\lvert 2.0y \rvert$
P-Net-NPT*	1.21 ± 0.94	3.73	NA	NA
P-Net-NPT$^{*}_{bs16-fil8-incr2}$	1.04 ± 0.80	3.83	NA	NA
P-Net-PT$^{*}_{bs16-mae}$	1.04 ± 0.83	3.92	NA	NA
P-Net-PT$^{*}_{bs16-mse}$	1.22 ± 0.85	4.17	NA	NA
P-Net-PT$^{*}_{bs16-mse-lr0.0001}$	1.14 ± 0.86	4.25	NA	NA
P-Net-PT$^{*}_{bs64-rmse}$	1.18 ± 0.89	4.35	NA	NA
M2-NPT-NOSEGM	0.97 ± 0.84	4.76	62.22	88.47
M2-NPT$_{lr0.001}$	0.86 ± 0.65	3.37	65.88	93.71
M2-NPT$_{lr0.0005}$	0.86 ± 0.64	3.41	66.07	93.86
M2-PT$_{lr0.0001}$	0.82 ± 0.64	3.09	68.69	93.43
M3-NPT	0.84 ± 0.68	3.29	67.95	92.24
M2-3D-PT**	0.81 ± 0.65	2.69	69.71	93.14

pt/*npt*: pretrained/not-pretrained CNN with weights from segmentation task
nosegm: standardized image VOIs were used as inputs to the CNN without masking
$*$: CNN from [148] but trained on a larger dataset an evaluated over 30 training rounds
$**$: 3D CNN produces one age prediction per MRI → table shows final prediction of the test set (no improvement through ML algorithms possible)

Table E.6: Performance of other age regression models using sagittal MRIs. Metrics were averaged over multiple training rounds of a single fold.

Model	MAE ± SD	Max AE	% ≤ $\lvert 1.0y \rvert$	% ≤ $\lvert 2.0y \rvert$
M2-NPT-NOSEGM	0.99 ± 0.82	6.10	60.56	89.57
M2-PT	0.91 ± 0.76	5.08	64.87	91.73
M2-3D-PT$^{*}_{lr0.0001}$	0.97 ± 0.75	3.57	64.00	90.67
M2-3D-PT$^{*}_{lr0.001}$	0.92 ± 0.74	3.68	62.93	92.00

pt/*npt*: pretrained/not-pretrained CNN with weights from segmentation task
nosegm: standardized image VOIs were used as inputs to the CNN without masking them with the segmentation maps
$*$: 3D CNN produces one age prediction per MRI → table shows final prediction of the test set (no improvement through ML algorithms possible)

Table E.7: Performance of multiple M1 models on *majority classification* in an "extended" 5-fold cross-validation using anthropometric measurements (AM) and ossification stages (OS). Metrics were computed for the 18-year-limit and averaged over multiple training rounds and five folds.

Model	Acc. ± SD	Sens. ± SD	Spec. ± SD	AUC ± SD
stat*	61.33 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	50.00 ± 0.00
M1-AM-KNC$_{def}$	70.29 ± 7.58	76.36 ± 7.82	60.00 ± 11.31	73.81 ± 8.26
M1-AM-RFC$_{def}$	69.09 ± 8.13	72.42 ± 10.28	63.44 ± 16.48	72.45 ± 2.10
M1-AM-KNC$_{opt}$	77.71 ± 2.14	80.00 ± 2.23	73.85 ± 3.77	76.92 ± 2.34
M1-AM-SVC$_{opt}$	76.00 ± 5.60	75.46 ± 9.79	76.92 ± 13.76	77.90 ± 9.63
M1-OS-GBC$_{def}$	80.57 ± 6.10	81.82 ± 7.04	78.46 ± 7.54	83.36 ± 5.52
M1-OS-RFC$_{def}$	77.62 ± 8.85	76.79 ± 11.90	79.03 ± 9.13	82.87 ± 4.89
M1-OS-GBC$_{opt}$	81.14 ± 6.41	82.73 ± 7.82	78.46 ± 7.54	83.18 ± 5.74
M1-OS-KNC$_{opt}$	80.57 ± 6.62	81.82 ± 10.76	78.46 ± 7.54	84.55 ± 6.43
M1-SKJ-GBC$_{def}$	80.00 ± 7.45	80.91 ± 9.27	78.46 ± 5.76	83.15 ± 7.05
M1-SKJ-KNC$_{def}$	80.00 ± 7.45	80.91 ± 9.27	78.46 ± 5.76	80.80 ± 8.03
M1-SKJ-GBC$_{opt}$	80.00 ± 7.45	80.91 ± 9.27	78.46 ± 5.76	83.15 ± 7.05
M1-SKJ-KNC$_{opt}$	80.00 ± 7.45	80.91 ± 9.27	78.46 ± 5.76	80.80 ± 8.03
M1-AM-SKJ-ETC$_{def}$	73.83 ± 8.50	75.91 ± 10.00	70.31 ± 13.74	82.38 ± 6.14
M1-AM-SKJ-LOG$_{def}$	74.29 ± 6.52	69.09 ± 7.82	83.08 ± 5.76	83.92 ± 8.81
M1-AM-SKJ-SVC$_{def}$	73.71 ± 7.32	68.18 ± 9.09	83.08 ± 5.76	85.46 ± 6.20
M1-AM-SKJ-ETC$_{opt}$	76.74 ± 7.64	71.55 ± 7.46	85.54 ± 9.27	85.87 ± 6.05
M1-AM-SKJ-LOG$_{opt}$	74.86 ± 6.62	70.00 ± 7.93	83.08 ± 5.76	84.06 ± 8.48
M1-AM-SKJ-SVC$_{opt}$	76.57 ± 5.24	70.00 ± 8.43	87.69 ± 9.23	85.56 ± 4.19

∗: static prediction all minors
def/opt machine learning algorithms with default or optimized parameters
OS/SKJ: ossification degree classification via three-stage system by Jopp et al. [92] or via *SKJ*

Table E.8: Performance of multiple M2 models on *majority classification* in an "extended" 5-fold cross-validation using *coronal* MRIs. Metrics were computed for the 18-year-limit and averaged over multiple training rounds and five folds.

Model	Acc. ± SD	Sens. ± SD	Spec. ± SD	AUC ± SD
stat*	49.25 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	50.00 ± 0.00
M2-KNC-NPT$_{all-def}$	81.54 ± 3.97	77.82 ± 6.21	87.85 ± 6.90	86.73 ± 3.58
M2-RFC-NPT$_{all-def}$	81.71 ± 4.75	78.77 ± 6.67	86.68 ± 7.40	88.15 ± 3.82
M2-SVC-NPT$_{all-def}$	81.94 ± 3.98	78.82 ± 5.50	87.23 ± 7.32	89.66 ± 3.36
M2-KNC$_{all-def}$	82.51 ± 2.78	79.73 ± 4.28	87.23 ± 6.46	86.41 ± 3.52
M2-RFC$_{all-def}$	82.89 ± 3.37	81.06 ± 5.05	86.00 ± 7.03	88.45 ± 3.27
M2-SVC$_{all-def}$	82.97 ± 2.62	80.18 ± 3.61	87.69 ± 6.34	89.82 ± 2.78
M2-ETC-FEATS$_{all-def}$	82.71 ± 3.12	79.75 ± 4.63	87.72 ± 6.49	88.34 ± 3.25
M2-RFC-FEATS$_{all-def}$	82.71 ± 3.30	79.66 ± 4.95	87.88 ± 6.72	88.56 ± 3.38
M2-SVC-FEATS$_{all-def}$	82.17 ± 3.64	78.36 ± 5.93	88.62 ± 6.20	89.25 ± 2.79
M2-KNC$_{all-opt}$	83.49 ± 2.16	80.91 ± 3.96	87.85 ± 6.17	87.22 ± 3.63
M2-RFC$_{all-opt}$	84.17 ± 3.19	83.09 ± 4.46	86.00 ± 7.33	89.91 ± 2.55
M2-SVC$_{all-opt}$	85.71 ± 4.00	86.36 ± 5.06	84.62 ± 6.53	90.82 ± 2.12
M2-RFC-FEATS$_{all-opt}$	83.49 ± 2.76	81.36 ± 4.38	87.08 ± 6.78	89.55 ± 2.77
M2-SVC-FEATS$_{all-opt}$	82.40 ± 3.67	83.82 ± 7.11	80.00 ± 12.96	89.50 ± 2.59
M2-KNC$_{best-def}$	85.14 ± 2.14	82.82 ± 2.88	90.77 ± 7.54	89.51 ± 3.18
M2-RFC$_{best-def}$	89.14 ± 2.14	89.09 ± 6.17	89.23 ± 7.85	92.52 ± 1.37
M2-SVC$_{best-def}$	85.14 ± 2.14	82.73 ± 3.40	89.23 ± 6.15	91.47 ± 3.07
M2-ETC-FEATS$_{best-def}$	88.57 ± 1.81	86.36 ± 4.07	92.31 ± 4.87	91.15 ± 2.14
M2-GBC-FEATS$_{best-def}$	87.43 ± 1.40	88.18 ± 4.64	86.15 ± 7.54	91.19 ± 2.72
M2-RFC-FEATS$_{best-def}$	89.71 ± 1.40	88.18 ± 3.64	92.31 ± 4.87	91.99 ± 3.04
M2-KNC$_{best-opt}$	86.29 ± 1.14	84.55 ± 3.64	89.23 ± 6.15	90.39 ± 3.58
M2-RFC$_{best-opt}$	88.00 ± 3.33	87.27 ± 4.45	89.23 ± 7.85	91.89 ± 2.14
M2-SVC$_{best-opt}$	89.14 ± 2.80	90.00 ± 3.40	87.69 ± 6.15	91.19 ± 2.11
M2-RFC-FEATS$_{best-opt}$	87.43 ± 2.29	86.36 ± 4.07	89.23 ± 6.15	91.71 ± 3.30
M2-SVC$_{best-opt}$	86.29 ± 2.14	87.27 ± 3.40	84.62 ± 6.88	90.07 ± 2.13

∗: static prediction all minors
npt: *not* pretrained on segmentation network weights
def/*opt*: machine learning algorithms with default or optimized parameters
all/*best*: all or best training rounds of each fold included

Table E.9: Performance of multiple M2 models on *majority classification* in an "extended" 5-fold cross-validation using *sagittal* MRIs. Metrics were computed for the 18-year-limit and averaged over multiple training rounds and five folds.

Model	Acc. ± SD	Sens. ± SD	Spec. ± SD	AUC ± SD
stat*	52.26 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	50.00 ± 0.00
M2-DTC$_{all-def}$	84.43 ± 3.96	82.71 ± 3.64	96.87 ± 7.91	84.79 ± 4.40
M2-ETC$_{all-def}$	87.81 ± 2.69	83.46 ± 2.83	93.99 ± 4.57	92.34 ± 1.95
M2-GBC$_{all-def}$	86.32 ± 3.28	83.16 ± 3.15	90.81 ± 6.78	92.13 ± 2.43
M2-KNC$_{all-def}$	87.73 ± 2.93	83.82 ± 3.07	93.29 ± 4.73	91.20 ± 2.43
M2-RFC$_{all-def}$	87.25 ± 2.97	83.63 ± 3.30	92.39 ± 5.56	92.22 ± 2.28
M2-SVC$_{all-def}$	88.27 ± 2.77	84.09 ± 2.88	94.19 ± 4.47	92.87 ± 1.69
M2-GBC$_{all-opt}$	86.91 ± 3.07	83.32 ± 2.71	92.00 ± 5.92	92.87 ± 2.50
M2-RFC$_{all-opt}$	87.63 ± 2.80	83.91 ± 2.76	92.90 ± 5.20	93.58 ± 2.01
M2-SVC$_{all-opt}$	87.47 ± 3.10	88.41 ± 2.24	86.13 ± 7.56	94.33 ± 1.45
M2-DTC$_{best-def}$	89.87 ± 2.00	87.27 ± 2.32	93.55 ± 4.56	90.41 ± 2.25
M2-ETC$_{best-def}$	90.93 ± 1.96	86.36 ± 1.44	97.42 ± 3.16	94.40 ± 1.57
M2-GBC$_{best-def}$	89.60 ± 1.96	86.82 ± 1.70	93.55 ± 4.56	94.05 ± 2.72
M2-KNC$_{best-def}$	89.60 ± 2.59	85.46 ± 2.32	95.48 ± 2.29	92.90 ± 1.26
M2-RFC$_{best-def}$	90.93 ± 1.31	88.64 ± 1.44	94.19 ± 1.29	94.38 ± 1.97
M2-SVC$_{best-def}$	89.87 ± 2.47	85.46 ± 1.82	96.13 ± 3.76	93.99 ± 1.44
M2-GBC$_{best-opt}$	89.33 ± 2.23	85.91 ± 1.70	94.19 ± 3.76	94.22 ± 2.49
M2-RFC$_{best-opt}$	89.33 ± 1.69	86.36 ± 1.44	93.55 ± 3.53	94.25 ± 2.33
M2-SVC$_{best-opt}$	89.87 ± 2.75	90.00 ± 2.32	89.68 ± 6.58	95.01 ± 1.21

∗: static prediction all minors
def/*opt*: machine learning algorithms with default or optimized parameters
all/*best*: all or best training rounds of each fold included

Table E.10: Performance of multiple M3 models on *majority classification* in an "extended" 5-fold cross-validation using *coronal* MRIs. Metrics were computed for the 18-year-limit and averaged over multiple training rounds and five folds.

Model	Acc. ± SD	Sens. ± SD	Spec. ± SD	AUC ± SD
stat*	49.25 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	50.00 ± 0.00
M3-DTC$_{all-def}$	73.30 ± 7.20	74.72 ± 8.03	70.91 ± 13.55	72.81 ± 7.94
M3-ETC$_{all-def}$	74.71 ± 7.64	73.14 ± 8.10	77.37 ± 12.24	81.74 ± 8.49
M3-GBC$_{all-def}$	75.55 ± 8.07	75.95 ± 8.43	74.88 ± 12.94	82.74 ± 8.29
M3-KNC$_{all-def}$	74.00 ± 7.23	72.73 ± 7.98	76.15 ± 10.46	80.77 ± 7.03
M3-RFC$_{all-def}$	74.99 ± 7.84	73.27 ± 8.59	77.91 ± 11.49	82.25 ± 7.97
M3-SVM$_{all-def}$	76.40 ± 6.95	72.73 ± 8.53	82.62 ± 8.40	83.02 ± 8.31
M3-GBC$_{all-opt}$	76.34 ± 7.56	74.91 ± 8.19	78.77 ± 9.80	83.96 ± 7.28
M3-RFC$_{all-opt}$	76.23 ± 7.37	74.09 ± 8.97	79.85 ± 9.96	84.22 ± 8.14
M3-SVM$_{all-opt}$	75.94 ± 6.16	71.18 ± 8.26	84.00 ± 9.34	85.36 ± 7.41
M3-DTC$_{best-def}$	83.43 ± 5.54	87.27 ± 4.45	76.92 ± 10.88	82.10 ± 6.45
M3-ETC$_{best-def}$	85.14 ± 6.62	82.73 ± 10.91	89.23 ± 3.77	88.46 ± 4.44
M3-GBC$_{best-def}$	84.57 ± 7.36	87.27 ± 8.33	80.00 ± 10.43	89.30 ± 3.77
M3-KNC$_{best-def}$	81.14 ± 6.91	79.09 ± 8.43	84.62 ± 4.87	84.27 ± 4.68
M3-RFC$_{best-def}$	86.86 ± 4.98	85.46 ± 5.30	89.23 ± 11.51	88.53 ± 3.81
M3-SVM$_{best-def}$	82.29 ± 6.36	77.27 ± 11.13	90.77 ± 3.08	87.90 ± 5.14
M3-GBC$_{best-opt}$	81.71 ± 7.36	80.91 ± 6.68	83.08 ± 11.31	88.95 ± 4.36
M3-RFC$_{best-opt}$	82.29 ± 7.54	82.73 ± 8.81	81.54 ± 11.51	89.09 ± 5.16
M3-SVM$_{best-opt}$	82.86 ± 4.78	80.91 ± 9.71	86.15 ± 5.76	88.11 ± 6.12

*: static prediction all minors
def/opt: machine learning algorithms with default or optimized parameters
all/best: all or best training rounds of each fold included

Bibliography

[1] Abbassi, V. (1998). Growth and normal puberty. *Pediatrics*, 102:507–511.

[2] Ambellan, F., Tack, A., Ehlke, M., and Zachow, S. (2019). Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. *Medical image analysis*, 52:109–118.

[3] Auf der Mauer, M., Säring, D., Stanczus, B., Herrmann, J., Groth, M., and Jopp-van Well, E. (2018). A 2-year follow-up MRI study for the evaluation of an age estimation method based on knee bone development. *International Journal of Legal Medicine*, 133(1):205–215.

[4] Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

[5] Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B. (2009). PatchMatch: a randomized correspondence algorithm for structural image editing. In *ACM SIGGRAPH 2009 papers on - SIGGRAPH '09*. ACM Press.

[6] Baumann, P., Widek, T., Merkens, H., Boldt, J., Petrovic, A., Urschler, M., Kirnbauer, B., Jakse, N., and Scheurer, E. (2015). Dental age estimation of living persons: Comparison of MRI with OPG. *Forensic Science International*, 253:76–80.

[7] Belaroussi, B., Milles, J., Carme, S., Zhu, Y. M., and Benoit-Cattin, H. (2006). Intensity non-uniformity correction in MRI: Existing methods and their validation. *Medical Image Analysis*, 10(2):234–246.

[8] BrainSuite (2019). Bias field correction. `http://brainsuite.org/processing/surfaceextraction/bfc/`. Accessed: 2019-10-21.

[9] Brechbühler, C., Gerig, G., and Székely, G. (1996). Compensation of spatial inhomogeneity in MRI based on a parametric bias estimate. In *Lecture Notes in Computer Science*, pages 141–146. Springer Berlin Heidelberg.

[10] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[11] Briechle, K. and Hanebeck, U. D. (2001). Template matching using fast normalized cross correlation. In Casasent, D. P. and Chao, T.-H., editors, *Optical Pattern Recognition XII*. SPIE.

[12] Britting-Reimer, E. (2015). Altersbestimmung in Deutschland und im Europäischen Vergleich. *Jugendhilfe*.

[13] Buhmann, M. D., Melville, P., Sindhwani, V., Quadrianto, N., Buntine, W. L., Torgo, L., Zhang, X., Stone, P., Struyf, J., Blockeel, H., Driessens, K., Miikkulainen, R., Wiewiora, E., Peters, J., Tedrake, R., Roy, N., Morimoto, J., Flach, P. A., and Fürnkranz, J. (2011). Random Decision Forests. In *Encyclopedia of Machine Learning*, pages 827–827. Springer US.

[14] Bundesamt für Migration und Flüchtlinge (2017). Das Bundesamt in Zahlen 2016 - Asyl, Migration und Integration. `https://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Broschueren/bundesamt-in-zahlen-2016.pdf?__blob=publicationFile`. Accessed: 2019-10-21.

[15] BWH and 3D Slicer contributors (2019). 3D Slicer: a multi-platform, free and open source software package for visualization and medical image computing. `https://www.slicer.org/`. Accessed: 2019-10-21.

[16] Cameriere, R., Cingolani, M., Giuliodori, A., Luca, S. D., and Ferrante, L. (2012). Radiographic analysis of epiphyseal fusion at knee joint to assess likelihood of having attained 18 years of age. *International Journal of Legal Medicine*, 126(6):889–899.

[17] Cardoso, H. F. V. (2007). Environmental effects on skeletal versus dental development: Using a documented subadult skeletal sample to test a basic assumption in human osteological research. *American Journal of Physical Anthropology*, 132(2):223–233.

[18] Chen, J., Yang, L., Zhang, Y., Alber, M., and Chen, D. Z. (2016). Combining Fully Convolutional and Recurrent Neural Networks for 3D Biomedical Image Segmentation. *arXiv preprint arXiv:1609.01006v2*.

[19] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets,

Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848.

[20] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587v3*.

[21] Chlebus, G., Meine, H., Moltz, J. H., and Schenk, A. (2017). Neural Network-Based Automatic Liver Tumor Segmentation With Random Forest-Based Candidate Filtering. *arXiv preprint arXiv:1706.00842v3*.

[22] Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348v2*.

[23] Chollet, F. (2017a). *Deep Learning with Python*. Manning.

[24] Chollet, F. (2017b). Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[25] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432. Springer International Publishing.

[26] Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2015). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv preprint arXiv:1511.07289v5*.

[27] Cole, T. J., Rousham, E. K., Hawley, N. L., Cameron, N., Norris, S. A., and Pettifor, J. M. (2014). Ethnic and sex differences in skeletal maturation among the Birth to Twenty cohort in South Africa. *Archives of Disease in Childhood*, 100(2):138–143.

[28] Craig, J. G., Cody, D. D., and van Holsbeeck, M. (2004). The distal femoral and proximal tibial growth plates: MR imaging, three-dimensional modeling and estimation of area and volume. *Skeletal Radiology*, 33(6):337–344.

[29] Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random Forests. In *Ensemble Machine Learning*, pages 157–175. Springer US.

[30] Dam, E. B., Lillholm, M., Marques, J., and Nielsen, M. (2015). Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative. *Journal of Medical Imaging*, 2(2):024001.

[31] De Tobel, J., Hillewig, E., Bogaert, S., Deblaere, K., and Verstraete, K. (2016a). Magnetic resonance imaging of third molars: developing a protocol suitable for forensic age estimation. *Annals of Human Biology*, 44(2):130–139.

[32] De Tobel, J., Hillewig, E., and Verstraete, K. (2016b). Forensic age estimation based on magnetic resonance imaging of third molars: converting 2D staging into 3D staging. *Annals of Human Biology*, 44(2):121–129.

[33] De Tobel, J., Parmentier, G. I. L., Phlypo, I., Descamps, B., Neyt, S., Van De Velde, W. L., Politis, C., Verstraete, K. L., and Thevissen, P. W. (2018). Magnetic resonance imaging of third molars in forensic age estimation: comparison of the Ghent and Graz protocols focusing on apical closure. *International Journal of Legal Medicine*, 133(2):583–592.

[34] De Tobel, J., Phlypo, I., Fieuws, S., Politis, C., Verstraete, K. L., and Thevissen, P. W. (2017). Forensic age estimation based on development of third molars: a staging technique for magnetic resonance imaging. *The Journal of forensic odonto-stomatology*, 35:117–140.

[35] Dedouit, F., Auriol, J., Rousseau, H., Rougé, D., Crubézy, E., and Telmon, N. (2012). Age assessment by magnetic resonance imaging of the knee: A preliminary study. *Forensic Science International*, 217(1–3):232.e1–232.e7.

[36] Demirjian, A., Goldstein, H., and Tanner, J. M. (1973). A New System of Dental Age Assessment. *Human Biology*, 45(2):211–227.

[37] Deniz, C. M., Xiang, S., Hallyburton, S., Welbeck, A., Babb, J. S., Honig, S., Cho, K., and Chang, G. (2017). Segmentation of the Proximal Femur from MR Images using Deep Convolutional Neural Networks. *Scientific Reports*, 8.

[38] DICOM (2019). About DICOM. `https://www.dicomstandard.org/about/`. Accessed: 2019-10-21.

[39] Dodin, P., Martel-Pelletier, J., Pelletier, J.-P., and Abram, F. (2011). A fully automated human knee 3D MRI bone segmentation using the ray casting technique. *Medical & Biological Engineering & Computing*, 49(12):1413–1424.

[40] Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017). Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks. In *Communications in Computer and Information Science*, pages 506–517. Springer International Publishing.

[41] Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., and Brox, T. (2016). Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9):1734–1747.

[42] Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V. C. T., Shi, L., and Heng, P.-A. (2016). Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 35(5):1182–1195.

[43] Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., and Heng, P.-A. (2017). 3D deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, 41:40–54.

[44] Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). The Importance of Skip Connections in Biomedical Image Segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer International Publishing.

[45] Duffy, B. A., Zhang, W., Tang, H., Zhao, L., Law, M., Toga, A. W., and Kim, H. (2018). Retrospective correction of motion artifact affected structural MRI images using deep learning of simulated motion. In *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*.

[46] Dvorak, J. (2009). Detecting over-age players using wrist MRI: science partnering with sport to ensure fair play. *British Journal of Sports Medicine*, 43(12):884–885.

[47] Dvorak, J., George, J., Junge, A., and Hodler, J. (2006). Age determination by magnetic resonance imaging of the wrist in adolescent male football players. *British Journal of Sports Medicine*, 41(1):45–52.

[48] Dvorak, J., George, J., Junge, A., and Hodler, J. (2007). Application of MRI of the wrist for age determination in international U-17 soccer competitions. *British Journal of Sports Medicine*, 41(8):497–500.

[49] Eikvil, L., Kvaal, S. I., Teigland, A., Haugen, M., and Grøgaard, J. (2012). Age estimation in youths and young adults. *Norwegian Computing Center*.

[50] Ekizoglu, O., Hocaoglu, E., Can, I. O., Inci, E., Aksoy, S., and Bilgili, M. G. (2015a). Magnetic resonance imaging of distal tibia and calcaneus for forensic age estimation in living individuals. *International Journal of Legal Medicine*, 129(4):825–831.

[51] Ekizoglu, O., Hocaoglu, E., Can, I. O., Inci, E., Aksoy, S., and Sayin, I. (2015b). Spheno-occipital synchondrosis fusion degree as a method to estimate age: a preliminary, magnetic resonance imaging study. *Australian Journal of Forensic Sciences*, 48(2):159–170.

[52] Ekizoglu, O., Hocaoglu, E., Inci, E., Can, I. O., Aksoy, S., and Kazimoglu, C. (2016). Forensic age estimation via 3-T magnetic resonance imaging of ossification of the proximal tibial and distal femoral epiphyses: Use of a T2-weighted fast spin-echo technique. *Forensic Science International*, 260:102.e1–102.e7.

[53] Elster, A. D. (2019). MR Artifacts. `http://mriquestions.com/hellipmr-artifacts.html`. Accessed: 2019-10-21.

[54] Erasmus, L. J., Hurter, D., Naude, M., Kritzinger, H. G., and Acho, S. (2004). A short overview of MRI artefacts. *South African Journal of Radiology*, 8(2):13.

[55] Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

[56] European Asylum Support Office (2014). *EASO age assessment practice in Europe*. Publications Office, Luxembourg.

[57] European Asylum Support Office (2018). *EASO practical guide on age assessment : second edition*. Publications Office of the European Union, Luxembourg.

[58] Fan, F., Zhang, K., Peng, Z., hui Cui, J., Hu, N., and hua Deng, Z. (2016). Forensic age estimation of living persons from the knee: Comparison of MRI with radiographs. *Forensic Science International*, 268:145–150.

[59] Fantini, I., Rittner, L., Yasuda, C., and Lotufo, R. (2018). Automatic detection of motion artifacts on MRI using Deep CNN. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE.

[60] Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.-C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J. V., Pieper, S., and Kikinis, R. (2012). 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magnetic Resonance Imaging*, 30(9):1323–1341.

[61] Feltz, V. (2015). Age assessment for unaccompanied minors. *Mèdecins du monde International Network*.

[62] Fleischhauer, J. (2018). Minderjährige Flüchtlinge: Sind so kleine Hände. *Spiegel Online*.

[63] Friederichs, H. (2015). Alles nur grobe Schätzungen? *Zeit Online*.

[64] Fripp, J., Crozier, S., Warfield, S. K., and Ourselin, S. (2010). Automatic Segmentation and Quantitative Analysis of the Articular Cartilages From Magnetic Resonance Images of the Knee. *IEEE Transactions on Medical Imaging*, 29(1):55–64.

[65] Galić, I., Mihanović, F., Giuliodori, A., Conforti, F., Cingolani, M., and Cameriere, R. (2016). Accuracy of scoring of the epiphyses at the knee joint (SKJ) for assessing legal adult age of 18 years. *International Journal of Legal Medicine*, 130(4):1129–1142.

[66] George, J., Nagendran, J., and Azmi, K. (2010). Comparison study of growth plate fusion using MRI versus plain radiographs as used in age determination for exclusion of overaged football players. *British Journal of Sports Medicine*, 46(4):273–278.

[67] Gilsanz, V. and Ratib, O. (2011). *Hand bone age: a digital atlas of skeletal maturity*. Springer-Verlag GmbH.

[68] Gohlke, B. and Wölfle, J. (2009). Growth and Puberty in German Children. *Deutsches Aerzteblatt Online*.

[69] Goodfellow, I., Bengio, Y., and Courville, A. (2017). *Deep Learning*. The MIT Press.

[70] Gray, H. and Lewis, W. H. (1918). *Anatomy of the human body*. Lea & Febiger, Philadelphia and New York, twentieth edition.

[71] Greulich, W. W. and Pyle, S. I. (1959). Radiographic atlas of skeletal development of the hand and wrist. *The American Journal Of The Medical Sciences*, 238(3):393.

[72] Guo, Y., Olze, A., Ottow, C., Schmidt, S., Schulz, R., Heindel, W., Pfeiffer, H., Vieth, V., and Schmeling, A. (2015). Dental age estimation in living individuals using 3.0 T MRI of lower third molars. *International Journal of Legal Medicine*, 129(6):1265–1270.

[73] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. (2017). Brain tumor segmentation with Deep Neural Networks. *Medical Image Analysis*, 35:18–31.

[74] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv preprint arXiv:1502.01852v1*.

[75] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[76] Heimann, T. and Meinzer, H.-P. (2009). Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis*, 13(4):543–563.

[77] Hermanussen, M. (2013). *Auxology: Studying Human Growth and Development*. Schweizerbart Sche Vlgsb.

[78] Hermanussen, M., Lieberman, L. S., Janewa, V. S., Scheffler, C., Ghosh, A., Bogin, B., Godina, E., Kaczmarek, M., El-Shabrawi, M., Salama, E. E., Rühli, F. J., Staub, K., Woitek, U., Blaha, P., Aßmann, C., van Buuren, S., Lehmann, A., Satake, T., Thodberg, H. H., Jopp, E., Kirchengast, S., Tutkuviene, J., McIntyre, M. H., Wittwer-Backofen, U., Boldsen, J. L., Martin, D. D., and Meier, J. (2012). Diversity in auxology: between theory and practice Proceedings of the 18th Aschauer Soiree, 13th November 2010. *Anthropologischer Anzeiger*, 69(2):159–174.

[79] Hillewig, E., Degroote, J., der Paelt, T. V., Visscher, A., Vandemaele, P., Lutin, B., D'Hooghe, L., Vandriessche, V., Piette, M., and Verstraete, K. (2012). Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: towards more sound age estimates. *International Journal of Legal Medicine*, 127(3):677–689.

[80] Hillewig, E., Tobel, J. D., Cuche, O., Vandemaele, P., Piette, M., and Verstraete, K. (2010). Magnetic resonance imaging of the medial extremity of the clavicle in forensic bone age determination: a new four-minute approach. *European Radiology*, 21(4):757–767.

[81] Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116.

[82] Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. `https://ml.jku.at/publications/older/ch7.pdf`. Accessed: 2019-10-21.

[83] Hoffer, E., Hubara, I., and Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems 30*, pages 1729–1739.

[84] Hou, Z. (2006). A Review on MR Image Intensity Inhomogeneity Correction. *International Journal of Biomedical Imaging*, 2006:1–11.

[85] Iglesias, J. E. and Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219.

[86] Iglovikov, V. I., Rakhlin, A., Kalinin, A. A., and Shvets, A. A. (2018). Paediatric Bone Age Assessment Using Deep Convolutional Neural Networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 300–308. Springer International Publishing.

[87] Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*.

[88] Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE.

[89] Jiang, J., Trundle, P., and Ren, J. (2010). Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 34(8):617–631.

[90] Jopp, E. (2007). *Methoden zur Alters- und Geschlechtsbestimmung auf dem Prüfstand : eine rechtsmedizinische empirische Studie*. Kovac, Hamburg.

[91] Jopp, E. (2013). *Die Abschlussphase des menschlichen Wachstums Longitudinale Ganzkörper- und Unterschenkelmessungen (Knemometrie) an jungen Erwachsenen zur Bestimmung des biologischen Alters und für forensische Zwecke*. Kovac, Hamburg.

[92] Jopp, E., Schröder, I., Maas, R., Adam, G., and Püschel, K. (2010). Proximale Tibiaepiphyse im Magnetresonanztomogramm. *Rechtsmedizin*, 20(6):464–468.

[93] Jopp, E., Schröder, I., Püschel, K., and Hermanussen, M. (2012). Longitudinal shrinkage in lower legs: "Negative growth" in healthy late-adolescent males. *Anthropologischer Anzeiger*, 69(1):107–115.

[94] Juntu, J., Sijbers, J., Dyck, D., and Gielen, J. (2005). Bias Field Correction for MRI Images. In *Advances in Soft Computing*, pages 543–551. Springer Berlin Heidelberg.

[95] Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78.

[96] Kayalibay, B., Jensen, G., and van der Smagt, P. (2017). CNN-based Segmentation of Medical Imaging Data. *arXiv preprint arXiv:1701.03056v2*.

[97] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv preprint arXiv:1609.04836v2*.

[98] Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980v9*.

[99] Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., and Biller, A. (2016). Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, 129:460–469.

[100] Knoll, F., Hammernik, K., Kobler, E., Pock, T., Recht, M. P., and Sodickson, D. K. (2018). Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic Resonance in Medicine*, 81(1):116–128.

[101] Knußmann, R. (1988). *Anthropologie: Handbuch der vergleichenden Biologie des Menschen*, volume 1, chapter Somatometrie, pages 232–285. Gustav Fischer.

[102] Koch, B. (2006). *Untersuchungen zur Anwendbarkeit der Skelettalterbestimmungsmethoden nach Greulich und Pyle sowie Thiemann und Nitz in der forensischen Altersdiagnostik bei Lebenden*. PhD thesis, Freie Universität Berlin.

[103] Köhler, S., Schmelzte, R., Loitz, C., and Püschel, K. (1994). Die Entwicklung des Weisheitszahnes als Kriterium der Lebensaltersbestimmung. *Annals of Anatomy - Anatomischer Anzeiger*, 176(4):339–345.

[104] Krämer, J. A., Schmidt, S., Jürgens, K.-U., Lentschig, M., Schmeling, A., and Vieth, V. (2014a). Forensic age estimation in living individuals using 3.0T MRI of the distal femur. *International Journal of Legal Medicine*, 128(3):509–514.

[105] Krämer, J. A., Schmidt, S., Jürgens, K.-U., Lentschig, M., Schmeling, A., and Vieth, V. (2014b). The use of magnetic resonance imaging to examine ossification of the proximal tibial epiphysis for forensic age estimation in living individuals. *Forensic Science, Medicine, and Pathology*, 10(3):306–313.

[106] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*, pages 1097–1105.

[107] Krupa, K. and Bekiesinska-Figatowska, M. (2015). Artifacts in magnetic resonance imaging. *Polish journal of radiology*, 80:93–106.

[108] Kubilay, S. (2016). Ablauf des deutschen Asylverfahrens. `https://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Broschueren/das-deutsche-asylverfahren.html`. Accessed: 2019-10-21.

[109] Laor, T., Chun, G. F. H., Dardzinski, B. J., Bean, J. A., and Witte, D. P. (2002). Posterior Distal Femoral and Proximal Tibial Metaphyseal Stripes at MR Imaging in Children and Young Adults. *Radiology*, 224(3):669–674.

[110] Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., and Langlotz, C. P. (2018). Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Radiology*, 287(1):313–322.

[111] Lasko, T. A., Bhagwat, J. G., Zou, K. H., and Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38(5):404–415.

[112] Lathuilière, S., Mesejo, P., Alameda-Pineda, X., and Horaud, R. (2018). A Comprehensive Analysis of Deep Regression. *arXiv preprint arXiv:1803.08450v2*.

[113] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1990). Handwritten Digit Recognition with a Back-Propagation Network. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan-Kaufmann.

[114] LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient BackProp. In *Lecture Notes in Computer Science*, pages 9–48. Springer Berlin Heidelberg.

[115] Lee, H., Tajmir, S., Lee, J., Zissen, M., Yeshiwas, B. A., Alkasab, T. K., Choy, G., and Do, S. (2017). Fully Automated Deep Learning System for Bone Age Assessment. *Journal of Digital Imaging*, 30(4):427–441.

[116] Lehmann, G. (2007). Label object representation and manipulation with ITK. *Insight Journal*.

[117] Lemley, J., Bazrafkan, S., and Corcoran, P. (2017). Smart Augmentation Learning an Optimal Data Augmentation Strategy. *IEEE Access*, 5:5858–5869.

[118] Lenz, P.-L. (2014). Altersbestimmung bei Flüchtlingen: Schlecht geschätzt. *Spiegel Online*.

[119] Li, Y., Huang, Z., Dong, X., Liang, W., Xue, H., Zhang, L., Zhang, Y., and Deng, Z. (2018). Forensic age estimation for pelvic X-ray images using deep learning. *European Radiology*, 29(5):2322–2329.

[120] Lin, G., Milan, A., Shen, C., and Reid, I. (2016). RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. *arXiv preprint arXiv:1611.06612v3*.

[121] Lin, M., Chen, Q., and Yan, S. (2013). Network In Network. *arXiv preprint arXiv:1312.4400v3*.

[122] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

[123] Lockemann, U., Fuhrmann, A., Püschel, K., Schmeling, A., and Geserick, G. (2004). Arbeitsgemeinschaft für Forensische Altersdiagnostik der Deutschen Gesellschaft für Rechtsmedizin. *Rechtsmedizin*, 14(2):123–126.

[124] Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30 of *1*, page 3.

[125] Mansour, H., Fuhrmann, A., Paradowski, I., Jopp-van Well, E., and Püschel, K. (2016). The role of forensic medicine and forensic dentistry in estimating the chronological age of living individuals in Hamburg, Germany. *International Journal of Legal Medicine*, 131(2):593–601.

[126] Mardani, M., Gong, E., Cheng, J. Y., Vasanawala, S. S., Zaharchuk, G., Xing, L., and Pauly, J. M. (2019). Deep Generative Adversarial Neural Networks for Compressive Sensing MRI. *IEEE Transactions on Medical Imaging*, 38(1):167–179.

[127] Marshall, W. A. and Tanner, J. M. (1969). Variations in pattern of pubertal changes in girls. *Archives of disease in childhood*, 44(235):291.

[128] Marshall, W. A. and Tanner, J. M. (1970). Variations in the pattern of pubertal changes in boys. *Archives of disease in childhood*, 45(239):13–23.

[129] Martin, R. (1914). *Lehrbuch der Anthropologie in systematischer Darstellung: mit besonderer Berücksichtigung der anthropologischen Methoden für Studierende Ärzte und Forschungsreisende*. Gustav Fischer.

[130] Mellin, W. D. (1957). Work with new electronic 'brains' opens field for army math experts. *The Hammond Times*, 10:66.

[131] MeVis Medical Solutions AG and Fraunhofer MEVIS (2019). MeVisLab. `https://www.mevislab.de/`. Accessed: 2019-10-21.

[132] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE.

[133] Mincer, H. H., Harris, E. F., and Berryman, H. E. (1993). The A.B.F.O. Study of Third Molar Development and Its Use as an Estimator of Chronological Age. *Journal of Forensic Sciences*, 38(2):13418J.

[134] Moeskops, P., Veta, M., Lafarge, M. W., Eppenhof, K. A. J., and Pluim, J. P. W. (2017). Adversarial Training and Dilated Convolutions for Brain MRI Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 56–64. Springer International Publishing.

[135] Mora, S., Boechat, M. I., Pietka, E., Huang, H. K., and Gilsanz, V. (2001). Skeletal Age Determinations in Children of European and African Descent: Applicability of the Greulich and Pyle Standards. *Pediatric Research*, 50(5):624–628.

[136] Mostad, P. and Tamsen, F. (2018). Error rates for unvalidated medical age assessment procedures. *International Journal of Legal Medicine*, 133(2):613–623.

[137] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[138] Naz, S., Majeed, H., and Irshad, H. (2010). Image segmentation using fuzzy clustering: A survey. In *2010 6th International Conference on Emerging Technologies (ICET)*. IEEE.

[139] Nekrasov, V., Ju, J., and Choi, J. (2016). Global Deconvolutional Networks for Semantic Segmentation. *arXiv preprint arXiv:1602.03930v2*.

[140] Neumayer, B., Schloegl, M., Payer, C., Widek, T., Tschauner, S., Ehammer, T., Stollberger, R., and Urschler, M. (2018). Reducing acquisition time for MRI-based forensic age estimation. *Scientific Reports*, 8(1).

[141] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S. A., de Marvao, A., Dawes, T., O'Regan, D. P., Kainz, B., Glocker, B., and Rueckert, D. (2018). Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation. *IEEE Transactions on Medical Imaging*, 37(2):384–395.

[142] Ontell, F. K., Ivanovic, M., Ablin, D. S., and Barlow, T. W. (1996). Bone age in children of diverse ethnicity. *American Journal of Roentgenology*, 167(6):1395–1398.

[143] Ottow, C., Schulz, R., Pfeiffer, H., Heindel, W., Schmeling, A., and Vieth, V. (2017). Forensic age estimation by magnetic resonance imaging of the knee: the definite relevance in bony fusion of the distal femoral- and the proximal tibial epiphyses using closest-to-bone T1 TSE sequence. *European Radiology*, 27(12):5041–5048.

[144] Patil, D. D. and Deore, S. G. (2013). Medical image segmentation: a review. *International Journal of Computer Science and Mobile Computing*, 2(1):22–27.

[145] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12:2825–2830.

[146] Pelt, D. M. and Sethian, J. A. (2017). A mixed-scale dense convolutional neural network for image analysis. *Proceedings of the National Academy of Sciences*, 115(2):254–259.

[147] Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., and Nielsen, M. (2013). Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network. In *Advanced Information Systems Engineering*, pages 246–253. Springer Berlin Heidelberg.

[148] Pröve, P.-L., Jopp-van Well, E., Stanczus, B., Morlock, M. M., Herrmann, J., Groth, M., Säring, D., and Auf der Mauer, M. (2018). Automated segmentation of the knee for age assessment in 3D MR images using convolutional neural networks. *International Journal of Legal Medicine*, 133(4):1191–1205.

[149] Qin, C., Schlemper, J., Caballero, J., Price, A. N., Hajnal, J. V., and Rueckert, D. (2019). Convolutional Recurrent Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 38(1):280–290.

[150] Quirmbach, F., Ramsthaler, F., and Verhoff, M. A. (2009). Evaluation of the ossification of the medial clavicular epiphysis with a digital ultrasonic system to

determine the age threshold of 21 years. *International Journal of Legal Medicine*, 123(3):241–245.

[151] Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., and Vaidya, V. (2016). Understanding the Mechanisms of Deep Transfer Learning for Medical Images. In *Deep Learning and Data Labeling for Medical Applications*, pages 188–196. Springer International Publishing.

[152] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[153] Redmon, J. and Farhadi, A. (2017). YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[154] Redmon, J. and Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767v1*.

[155] Reisinger, W., Kleiber, M., Beckmann, D., Buhr, J., Erben, U., Erfurth, F., Filz, G.-R., Genssler, W., Harzheim, I., Illing, A., Illing, H., Joachim, H.-E., Klaer, U., Preuss, H.-J., Roick, H., Rupprecht, E., Schultz-Wernitz, C., Stoye, H.-D., Weingärtner, R., Brettschneider, I., and Ziegler, P.-F. (2006). Forensische Altersdiagnostik im Strafverfahren. In Thiemann, H.-H., Nitz, I., and Schmeling, A., editors, *Röntgenatlas der normalen Hand im Kindesalter*. Georg Thieme Verlag.

[156] Roche, A. F. (1979). Secular trends in human growth, maturation, and development. *Monographs of the Society for Research in Child Development*, 44:1–120.

[157] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing.

[158] Roth, H. R., Lee, C. T., Shin, H.-C., Seff, A., Kim, L., Yao, J., Lu, L., and Summers, R. M. (2015). Anatomy-specific classification of medical images using deep convolutional nets. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. IEEE.

[159] Roth, H. R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., and Mori, K. (2018). An application of cascaded 3D fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99.

[160] Saint-Martin, P., Rérolle, C., Dedouit, F., Bouilleau, L., Rousseau, H., Rougé, D., and Telmon, N. (2013). Age estimation by magnetic resonance imaging of the distal tibial epiphysis and the calcaneum. *International Journal of Legal Medicine*, 127(5):1023–1030.

[161] Saint-Martin, P., Rérolle, C., Dedouit, F., Rousseau, H., Rougé, D., and Telmon, N. (2014a). Evaluation of an automatic method for forensic age estimation by magnetic resonance imaging of the distal tibial epiphysis—a preliminary study focusing on the 18-year threshold. *International Journal of Legal Medicine*, 128(4):675–683.

[162] Saint-Martin, P., Rérolle, C., Pucheux, J., Dedouit, F., and Telmon, N. (2014b). Contribution of distal femur MRI to the determination of the 18-year limit in forensic age estimation. *International Journal of Legal Medicine*, 129(3):619–620.

[163] Salehi, S. S. M., Erdogmus, D., and Gholipour, A. (2017). Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In *Machine Learning in Medical Imaging*, pages 379–387. Springer International Publishing.

[164] Salter, R. B. and Harris, W. R. (1963). Injuries Involving the Epiphyseal Plate. *Journal of Bone and Joint Surgery*, 45(3):587–622.

[165] Säring, D., Auf der Mauer, M., and Jopp, E. (2014). Klassifikation des Verschlussgrades der Epiphyse der proximalen Tibia zur Altersbestimmung. In *Informatik aktuell*, pages 60–65. Springer Berlin Heidelberg.

[166] Scherer, D., Müller, A., and Behnke, S. (2010). Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. In *Artificial Neural Networks – ICANN 2010*, pages 92–101. Springer Berlin Heidelberg.

[167] Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., and Rueckert, D. (2018). A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 37(2):491–503.

[168] Schmeling, A. (2011). Forensische Altersdiagnostik bei lebenden Jugendlichen und jungen Erwachsenen. *Rechtsmedizin*, 21(2):151–162.

[169] Schmeling, A., Dettmeyer, R., Rudolf, E., Vieth, V., and Geserick, G. (2016). Forensic Age Estimation: Methods, Certainty, and the Law. *Deutsches Arzteblatt international*, 113:44–50.

[170] Schmeling, A., Grundmann, C., Fuhrmann, A., Kaatsch, H.-J., Knell, B., Ramsthaler, F., Reisinger, W., Riepert, T., Ritz-Timme, S., Rösing, F. W., Rötzscher, K., and Geserick, G. (2008). Aktualisierte Empfehlungen der Arbeitsgemeinschaft für Forensische Altersdiagnostik für Altersschätzungen bei Lebenden im Strafverfahren. *Rechtsmedizin*, 18(6):451–453.

[171] Schmeling, A., Kaatsch, H.-J., Marré, B., Reisinger, W., Riepert, T., Ritz-Timme, S., Rösing, F. W., Rötzscher, K., and Geserick, G. (2001). Empfehlungen für die Altersdiagnostik bei Lebenden im Strafverfahren. *Rechtsmedizin*, 11(1):1–3.

[172] Schmeling, A., Manuel, P., Luis, J., and Irene, M. (2011). Forensic Age Estimation in Unaccompanied Minors and Young Living Adults. In *Forensic Medicine - From Old Problems to New Challenges*. InTech.

[173] Schmeling, A., Olze, A., Reisinger, W., König, M., and Geserick, G. (2003). Statistical analysis and verification of forensic age estimation of living persons in the Institute of Legal Medicine of the Berlin University Hospital Charité. *Legal Medicine*, 5:S367–S371.

[174] Schmeling, A., Reisinger, W., Loreck, D., Vendura, K., Markus, W., and Geserick, G. (2000). Effects of ethnicity on skeletal maturation: consequences for forensic age estimations. *International Journal of Legal Medicine*, 113(5):253–258.

[175] Schmidt, S., Baumann, U., Schulz, R., Reisinger, W., and Schmeling, A. (2007a). Study of age dependence of epiphyseal ossification of the hand skeleton. *International Journal of Legal Medicine*, 122(1):51–54.

[176] Schmidt, S., Koch, B., Schulz, R., Reisinger, W., and Schmeling, A. (2007b). Comparative analysis of the applicability of the skeletal age determination methods of Greulich–Pyle and Thiemann–Nitz for forensic age estimation in living subjects. *International Journal of Legal Medicine*, 121(4):293–296.

[177] Schmidt, S., Mühler, M., Schmeling, A., Reisinger, W., and Schulz, R. (2007c). Magnetic resonance imaging of the clavicular ossification. *International Journal of Legal Medicine*, 121(4):321–324.

[178] Schmidt, S., Schiborr, M., Pfeiffer, H., Schmeling, A., and Schulz, R. (2013a). Age dependence of epiphyseal ossification of the distal radius in ultrasound diagnostics. *International Journal of Legal Medicine*, 127(4):831–838.

[179] Schmidt, S., Schiborr, M., Pfeiffer, H., Schmeling, A., and Schulz, R. (2013b). Sonographic examination of the apophysis of the iliac crest for forensic age estimation in living persons. *Science & Justice*, 53(4):395–401.

[180] Schmidt, S., Vieth, V., Timme, M., Dvorak, J., and Schmeling, A. (2015). Examination of ossification of the distal radial epiphysis using magnetic resonance imaging. New insights for age estimation in young footballers in FIFA tournaments. *Science & Justice*, 55(2):139–144.

[181] Schulz, R., Mühler, M., Reisinger, W., Schmidt, S., and Schmeling, A. (2007). Radiographic staging of ossification of the medial clavicular epiphysis. *International Journal of Legal Medicine*, 122(1):55–58.

[182] Schulz, R., Schiborr, M., Pfeiffer, H., Schmidt, S., and Schmeling, A. (2014a). Forensic age estimation in living subjects based on ultrasound examination of the ossification of the olecranon. *Journal of Forensic and Legal Medicine*, 22:68–72.

[183] Schulz, R., Schmidt, S., Pfeiffer, H., and Schmeling, A. (2014b). Sonographische Untersuchungen verschiedener Skelettregionen. *Rechtsmedizin*, 24(6):480–484.

[184] Seil, R., Frosch, K.-H., and Becker, R. (2012). Kreuzbandverletzungen im Wachstumsalter. *SFA Arthroskopie Aktuell*, 25.

[185] Serin, J., Rérolle, C., Pucheux, J., Dedouit, F., Telmon, N., Savall, F., and Saint-Martin, P. (2016). Contribution of magnetic resonance imaging of the wrist and hand to forensic age assessment. *International Journal of Legal Medicine*, 130(4):1121–1128.

[186] Serinelli, S., Panebianco, V., Martino, M., Battisti, S., Rodacki, K., Marinelli, E., Zaccagna, F., Semelka, R. C., and Tomei, E. (2015). Accuracy of MRI skeletal age estimation for subjects 12–19. Potential use for subjects of unknown age. *International Journal of Legal Medicine*, 129(3):609–617.

[187] Setiono, R. and Liu, H. (1997). Neural-network feature selector. *IEEE Transactions on Neural Networks*, 8(3):654–662.

[188] Shaw, R., Sudre, C., Ourselin, S., and Cardoso, M. J. (2019). MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty. In Cardoso, M. J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., and Vercauteren, T., editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 427–436, London, United Kingdom. PMLR.

[189] Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition*. IEEE Comput. Soc.

[190] Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97.

[191] Smith, T. and Brownlees, L. (2011). Age assessment practices: a literature review & annotated bibliography. Technical report, Child Protection Section, UNICEF.

[192] Son, J., Park, S. J., and Jung, K.-H. (2017). Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. *arXiv preprint arXiv:1706.09318v1*.

[193] Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., and Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. *Medical Image Analysis*, 36:41–51.

[194] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

[195] Stern, D., Ebner, T., Bischof, H., Grassegger, S., Ehammer, T., and Urschler, M. (2014). Fully Automatic Bone Age Estimation from Left Hand MR Images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, pages 220–227. Springer International Publishing.

[196] Štern, D., Kainz, P., Payer, C., and Urschler, M. (2017). Multi-factorial Age Estimation from Skeletal and Dental MRI Volumes. In *Machine Learning in Medical Imaging*, pages 61–69. Springer International Publishing.

[197] Stern, D., Payer, C., Giuliani, N., and Urschler, M. (2019). Automatic Age Estimation and Majority Age Classification From Multi-Factorial MRI Data. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1392–1403.

[198] Štern, D., Payer, C., and Urschler, M. (2019). Automated age estimation from MRI volumes of the hand. *Medical Image Analysis*, 58:101538.

[199] Stern, D. and Urschler, M. (2016). From individual hand bone age estimates to fully automated age estimation via learning-based information fusion. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE.

[200] Styner, M., Brechbuhler, C., Szckely, G., and Gerig, G. (2000). Parametric estimate of intensity inhomogeneities applied to MRI. *IEEE Transactions on Medical Imaging*, 19(3):153–165.

[201] Sun, S., Cao, Z., Zhu, H., and Zhao, J. (2019). A Survey of Optimization Methods from a Machine Learning Perspective. *arXiv preprint arXiv:1906.06821v2*.

[202] Tack, A., Mukhopadhyay, A., and Zachow, S. (2018). Knee menisci segmentation using convolutional neural networks: data from the Osteoarthritis Initiative. *Osteoarthritis and Cartilage*, 26(5):680–688.

[203] Tanner, J. M. (1983). *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. Academic Press, London New York.

[204] Tanner, J. M. and Davies, P. S. W. (1985). Clinical longitudinal standards for height and height velocity for North American children. *The Journal of Pediatrics*, 107(3):317–329.

[205] Terada, Y., Kono, S., Tamada, D., Uchiumi, T., Kose, K., Miyagi, R., Yamabe, E., and Yoshioka, H. (2012). Skeletal age assessment in children using an open compact MRI system. *Magnetic Resonance in Medicine*, 69(6):1697–1702.

[206] Terada, Y., Tamada, D., Kose, K., Nozaki, T., Kaneko, Y., Miyagi, R., and Yoshioka, H. (2015). Acceleration of skeletal age MR examination using compressed sensing. *Journal of Magnetic Resonance Imaging*, 44(1):204–211.

[207] Timme, M., Ottow, C., Schulz, R., Pfeiffer, H., Heindel, W., Vieth, V., Schmeling, A., and Schmidt, S. (2016). Magnetic resonance imaging of the distal radial epiphysis: a new criterion of maturity for determining whether the age of 18 has been completed? *International Journal of Legal Medicine*, 131(2):579–584.

[208] Tomei, E., Sartori, A., Nissman, D., Ansari, N. A., Battisti, S., Rubini, A., Stagnitti, A., Martino, M., Marini, M., Barbato, E., and Semelka, R. C. (2013). Value of MRI of the hand and the wrist in evaluation of bone age: Preliminary results. *Journal of Magnetic Resonance Imaging*, 39(5):1198–1205.

[209] Tscholl, P. M., Junge, A., Dvorak, J., and Zubler, V. (2015). MRI of the wrist is not recommended for age determination in female football players of U-16/U-17 competitions. *Scandinavian Journal of Medicine & Science in Sports*, 26(3):324–328.

[210] Tustison, N. and Gee, J. (2009). N4ITK: Nick's N3 ITK implementation for MRI bias field correction. *Insight Journal*, pages 1–9.

[211] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

[212] UNICEF (2011). *The State of the World's Children, 2011: Adolescence, an Age of Opportunity*. UNICEF.

[213] Urschler, M., Grassegger, S., and Štern, D. (2015). What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. *Annals of Human Biology*, 42(4):358–367.

[214] Vieth, V., Kellinghaus, M., Schulz, R., Pfeiffer, H., and Schmeling, A. (2010). Beurteilung des Ossifikationsstadiums der medialen Klavikulaepiphysenfuge. *Rechtsmedizin*, 20(6):483–488.

[215] Vieth, V., Schulz, R., Heindel, W., Pfeiffer, H., Buerke, B., Schmeling, A., and Ottow, C. (2018). Forensic age assessment by 3.0T MRI of the knee: proposal of a new MRI classification of ossification stages. *European Radiology*, 28(8):3255–3262.

[216] Vovk, U., Pernus, F., and Likar, B. (2007). A Review of Methods for Correction of Intensity Inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26(3):405–421.

[217] Vucic, S., de Vries, E., Eilers, P. H. C., Willemsen, S. P., Kuijpers, M. A. R., Prahl-Andersen, B., Jaddoe, V. W. V., Hofman, A., Wolvius, E. B., and Ongkosuwito, E. M. (2014). Secular trend of dental development in Dutch children. *American Journal of Physical Anthropology*, 155(1):91–98.

[218] Wang, S., Su, Z., Ying, L., Peng, X., Zhu, S., Liang, F., Feng, D., and Liang, D. (2016). Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE.

[219] Wang, Y., Song, Y., Xie, H., Li, W., Hu, B., and Yang, G. (2017). Reduction of Gibbs artifacts in magnetic resonance imaging based on Convolutional Neural Network. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE.

[220] Wittschieber, D., Ottow, C., Schulz, R., Püschel, K., Bajanowski, T., Ramsthaler, F., Pfeiffer, H., Vieth, V., Schmidt, S., and Schmeling, A. (2015). Forensic age diagnostics using projection radiography of the clavicle: a prospective multicenter validation study. *International Journal of Legal Medicine*, 130(1):213–219.

[221] Wolterink, J. M., Leiner, T., Viergever, M. A., and Išgum, I. (2017). Dilated Convolutional Neural Networks for Cardiovascular MR Segmentation in Congenital Heart Disease. In *Reconstruction, Segmentation, and Analysis of Medical Images*, pages 95–102. Springer International Publishing.

[222] Yang, Y., Sun, J., Li, H., and Xu, Z. (2017). ADMM-Net: A Deep Learning Approach for Compressive Sensing MRI. *arXiv preprint arXiv:1705.06869v1*.

[223] Yu, F. and Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv preprint arXiv:1511.07122v3*.

[224] Zaitsev, M., Maclaren, J., and Herbst, M. (2015). Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 42(4):887–901.

[225] Zbontar, J., Knoll, F., Sriram, A., Murrell, T., Huang, Z., Muckley, M. J., Defazio, A., Stern, R., Johnson, P., Bruno, M., Parente, M., Geras, K. J., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Yakubova, N., Pinkerton, J., Wang, D., Owens, E., Zitnick, C. L., Recht, M. P., Sodickson, D. K., and Lui, Y. W. (2018). fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *arXiv preprint arXiv:1811.08839v2*.

[226] Zeiler, M. D. and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing.

[227] Zhang, X., Zhou, X., Lin, M., and Sun, J. (2017). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *arXiv preprint arXiv:1707.01083v2*.

[228] Zhao, F., Huang, Q., and Gao, W. (2006). Image Matching by Normalized Cross-Correlation. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*. IEEE.

[229] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2016). Pyramid Scene Parsing Network. *arXiv preprint arXiv:1612.01105v2*.

[230] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Learning Deep Features for Discriminative Localization. *arXiv preprint arXiv:1512.04150v1*.