Data-Driven Classification Methods for Craniosynostosis Using 3D Surface Scans

Zur Erlangung des akademischen Grades eines

DOKTORS DER INGENIEURWISSENSCHAFTEN (Dr.-Ing.)

von der KIT-Fakultät für

Elektrotechnik und Informationstechnik

des Karlsruher Instituts für Technologie (KIT)

genehmigte

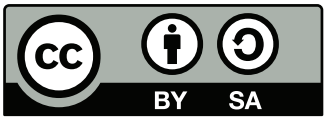
DISSERTATION

von

Matthias Schaufelberger, M.Sc.

geb. in Karlsruhe

Tag der mündlichen Prüfung:	14. Dezember 2023
Referent:	Prof. Dr. rer. nat. Werner Nahm
Korreferent:	Prof. Dr.-Ing. Michael Heizmann



Abstract

This work investigates into radiation-free classification of craniosynostosis with an additional focus on including data augmentation and using synthetic data as a replacement for clinical data.

Motivation: Craniosynostosis is a condition affecting infants and leads to head deformities. Diagnosis using radiation-free 3D surface scans is a promising alternative to traditional computed tomography (CT) imaging. Clinical data are only sparsely available due to the low prevalence and difficulties in anonymization. This work addresses these challenges by proposing new classification algorithms for craniosynostosis, by creating synthetic data for the scientific community, and by demonstrating that it is possible to fully replace clinical data with synthetic data without losing classification performance.

Methods: A statistical shape model (SSM) of craniosynostosis patients is created and made publicly available. A 3D-2D conversion from the 3D mesh geometry to a 2D image is proposed which enables the usage of convolutional neural networks (CNNs) and data augmentation in the image domain. Three classification approaches (based on cephalometric measurements, based on an SSM, and based on the 2D images using a CNN) to distinguish between three types of craniosynostosis and a control group are proposed and evaluated. Finally, the clinical training data are fully replaced with synthetic data by an SSM and a generative adversarial network (GAN).

Results: The proposed CNN classification outperformed competing approaches on a clinical dataset of 496 subjects and achieved an F1-score of 0.964. Data augmentation increased the F1-score to 0.975. Attribution maps of the classification decision showed high amplitudes on parts of the head associated with craniosynostosis. Replacing the clinical data with synthetic data created by an SSM and a GAN still yielded an F1-score of more than 0.95 without the model having seen a single clinical subject.

Conclusion: The proposed conversion of 3D geometry to a 2D encoded image improved performance to existing classifiers and enabled data augmentation during training. Using an SSM and a GAN, clinical training data could be replaced with synthetic data. This work improves existing diagnostic approaches on radiation-free recordings and demonstrates the usability of synthetic data which makes clinical applications more objective, interpretable, and less expensive.

Zusammenfassung

Diese Arbeit befasst sich mit strahlungsfreier Klassifizierung von Kraniosynostose mit zusätzlichem Schwerpunkt auf Datenaugmentierung und auf die Verwendung synthetischer Daten als Ersatz für klinische Daten.

Motivation: Kraniosynostose ist eine Erkrankung, die Säuglinge betrifft und zu Kopfdeformitäten führt. Diagnose mittels strahlungsfreier 3D Oberflächenscans ist eine vielversprechende Alternative zu traditioneller computertomographischer Bildgebung. Aufgrund der niedrigen Prävalenz und schwieriger Anonymisierbarkeit sind klinische Daten nur spärlich vorhanden. Diese Arbeit adressiert diese Herausforderungen, indem sie neue Klassifizierungsalgorithmen vorschlägt, synthetische Daten für die wissenschaftliche Gemeinschaft erstellt und zeigt, dass es möglich ist, klinische Daten vollständig durch synthetische Daten zu ersetzen, ohne die Klassifikationsleistung zu beeinträchtigen.

Methoden: Ein Statistisches Shape Modell (SSM) von Kraniosynostosepatienten wird erstellt und öffentlich zugänglich gemacht. Es wird eine 3D-2D-Konvertierung von der 3D-Gittergeometrie in ein 2D-Bild vorgeschlagen, die die Verwendung von Convolutional Neural Networks (CNNs) und Datenaugmentierung im Bildbereich ermöglicht. Drei Klassifizierungsansätze (basierend auf cephalometrischen Messungen, basierend auf dem SSM, und basierend auf den 2D Bildern mit einem CNN) zur Unterscheidung zwischen drei Pathologien und einer Kontrollgruppe werden vorgeschlagen und bewertet. Schließlich werden die klinischen Trainingsdaten vollständig durch synthetische Daten aus einem SSM und einem generativen adversarialen Netz (GAN) ersetzt.

Ergebnisse: Die vorgeschlagene CNN-Klassifikation übertraf konkurrierende Ansätze in einem klinischen Datensatz von 496 Probanden und erreichte einen F1-Score von 0,964. Datenaugmentierung erhöhte den F1-Score auf 0,975. Zuschreibungen der Klassifizierungsentscheidung zeigten hohe Amplituden an Teilen des Kopfes, die mit Kraniosynostose in Verbindung stehen. Das Ersetzen der klinischen Daten durch synthetische Daten, die mit einem SSM und einem GAN erstellt wurden, ergab noch immer einen F1-Score von über 0,95, ohne dass das Modell ein einziges klinisches Subjekt gesehen hatte.

Schlussfolgerung: Die vorgeschlagene Umwandlung von 3D-Geometrie in ein 2D-kodiertes Bild verbesserte die Leistung bestehender Klassifikatoren und ermöglichte eine Datenaugmentierung während des Trainings. Unter Verwendung eines SSM

und eines GANs konnten klinische Trainingsdaten durch synthetische Daten ersetzt werden. Diese Arbeit verbessert bestehende diagnostische Ansätze auf strahlungsfreien Aufnahmen und demonstriert die Verwendbarkeit von synthetischen Daten, was klinische Anwendungen objektiver, interpretierbarer, und weniger kostspielig machen.

Acknowledgments

First, I thank my supervisor Prof. Dr. rer. nat. Werner Nahm for accepting me as a PhD student and giving me the opportunity to work under his guidance, mentorship, support, and his omnipresent scientific curiosity and expertise. Second, I would also like to thank my second examinor Prof. Dr.-Ing. Michael Heizmann for co-refereeing this thesis and his sincere curiosity in my work.

Parts of this research were founded by the HEiKA Heidelberg Karlsruhe Strategic Partnership under grant "HEiKA_19–17". I am very grateful for the funding organization and fruitful collaboration with the project partners in Heidelberg. I am thankful for the scientific input, stimulating discussions and being provided with the clinical data. In particular, I would like to thank Prof. Dr. med. Dr. med. dent. Christian Freudlsperger and Dr. sc. hum. Urs Eisenmann for their clinical and technical expertise, Friedemann, Frederic, and Niclas for the collaboration, their feedback on various manuscripts and ideas, and especially Reinald who took care of all my clinical questions and helped me a lot in various stages of this scientific journey. A special thanks also goes to Andreas for his mentorship along the project.

Furthermore I would like to thank all my colleagues at the Institute of Biomedical Engineering (IBT) who created a great working atmosphere and made me less grumpy each day. A special thanks goes to the optics group, and especially to Alexander, Lorena, Lu, Miriam, Simon, and Tobias, who gave me valuable feedback and encouragement along the whole way and proofread parts of this thesis. Also thanks to Olaf, Francesca, and Axel for their critical and valuable feedback.

Thank you to all the students I supervised and who contributed to this project. I would especially like to highlight the work of my students Christian and Anna Maria.

Something which is often forgotten is the importance of free/libre and open source software for the research community, used by millions of people and scientists every day and often maintained and improved by thousands of volunteers. The community contributions often do not get enough credit, or are taken for granted. Representatively, I would like to thank GNU/Linux and the Vim text editor for boosting my productivity. I hope to give something back.

Finally, my deepest gratitude belongs to my family: I thank my parents Beatrice and Gerd and my sister Lea for their unconditional love and support in my endeav-

ors, and for having shaped me to the person I am now. Lastly, thank you Viola for your love, support and encouragement making my life brighter every single day.

Contents

Abbreviations

ADAM	adaptive moment estimation . . . . . 20 f., 90, 109
cGAN	conditional GAN 26
CI	cephalic index . . . . . 12 f., 59 ff., 64, 67 f., 127
CNN	convolutional neural network i, 5 f., 17, 21 f., 68, 79 f., 87 f., 91, 95, 97 f., 102, 107, 109, 112 f., 117, 119–122
CPD	coherent point drift 30, 123
CT	computed tomographyi, 1 f., 11 ff., 44, 60, 67, 69, 77 f., 97 f., 102, 118, 122
CVAI	cranial vault asymmetry index . . . . . 13, 59 ff., 64, 67 f., 127
DCGAN	deep convolutional generative adversarial network 26
DT	decision tree . . . . . 15 ff., 59 f., 62, 67, 73 f.
FNN	feedforward neural network 3, 17, 19 ff., 78 ff., 87, 91, 95, 97 f., 120 f.
FOSS	free and open-source software 15, 22
GAN	generative adversarial network i, 3, 5, 25 f., 102 f., 105 ff., 109 f., 112 f., 120 f., 137 f.
GPA	generalized Procrustes analysis 30, 32, 47
ICP	iterative closest points 30
ICPA	nonrigid iterative closest points affine 28, 30, 46, 49 f., 53, 124 f.
ICPD	iterative coherent point drift . . . . . 30, 46, 49 f., 123
ICPT	nonrigid iterative closest point translation . 30, 46, 49 f., 124 f.
kNN	k-nearest-neighbors . 16 f., 59 f., 62, 66 ff., 73 f., 76, 87, 119, 123
LBRP	Laplace-Beltrami regularized projection30 f., 46, 49 f., 53, 70, 123
LDA	linear discriminant analysis . 16 f., 20, 62, 66, 68, 73, 76 f., 131
MAP	maximum a posteriori 16
ML	machine learning . . . . . 1, 5, 15, 17, 20, 59, 68 f., 79, 119, 121
MLP	multi layer perceptron 20
MRI	magnetic resonance imaging 11, 98, 118
NB	naïve Bayes . . . . . 16, 62, 66, 73 f., 76 f., 87, 119, 131
NN	neural network . . . . . 3, 5, 15, 17, 20 f., 25, 91, 120
PCA	principal component analysis 3, 16, 28, 32 f., 47, 73 f., 78, 102 f., 105, 107, 109, 112 f., 120 f.
PDM	point distribution model 27

QDA	quadratic discriminant analysis 17
RANSAC	random sample consensus 28, 98
ReLU	rectified linear unit 19
RF	random forest . . . . . 16 f., 59 f., 62, 67, 73 f.
SGD	stochastic gradient descent . . . . . 20 f.
SHAP	SHapley Additive exPlanations . . . . . 62 f., 68, 87
SIDS	sudden infant death syndrome 12
$SSIM_{cc}$	structural similarity index measure to closest clinical sample 107, 109, 112
SSM	statistical shape modeli, 2 f., 5, 27, 32 f., 45–48, 50, 53, 55, 68–71, 73 f., 77 f., 80, 87, 90, 97, 101 ff., 105, 107, 109 f., 112 f., 117–121
STO	sellion tragion orientation 43, 60, 81
SVD	singular value decomposition 29, 47
SVM	support vector machine . . . . . 17, 62, 66 f., 73, 76, 131
WGAN	Wasserstein generative adversarial network 26
WPCA	weighted principal component analysis 32, 47, 71, 73

Chapter 1

Introduction

1.1 Motivation

In the past decade, machine learning (ML) approaches have been a driver for technological progress in biomedical engineering and clinical healthcare [1]. Their ability to learn an underlying function and extract statistical patterns from training data not observable by humans often make them more accurate, effective and less expensive compared to traditional approaches. Due to technological advancements in computing power and the trend toward collecting large amounts of data [2], data-driven approaches are likely to continue being one of the dominant drivers in the future of healthcare [3]. As such, they have been consistently advancing into image analysis, modeling, segmentation, and classification, especially if data are available in abundance. However, this is usually not the case for rare diseases such as craniosynostosis [4].

Craniosynostosis affects infants and is a condition characterized by irregular growth patterns of the skull due to the premature fusion of head sutures, leading to distinctive head deformities. Craniosynostosis is linked to increased intracranial pressure [5] which has been connected to reduced brain growth and diminished neurocognitive development in planning, attention, processing speed, vision, and speech [6, 7]. Surgical remodeling of the skull is performed to reduce intracranial pressure and to achieve a physiological head shape [8]. Early intervention and diagnosis are key to limit neurological consequences and to ensure regular skull growth [9]. Diagnosis requires visual assessment, palpation, and medical imaging with computed tomography (CT) being the gold standard [10]. CT imaging is not only expensive, but also exposes children to ionizing radiation which should be avoided and performed only when absolutely necessary [8]. In contrast, 3D photography such as stereophotography and laser scanning has been proposed for routinely monitoring and documenting patients [11]. This cost-effective and radiation-free image modality calls for the development of ML approaches to enable an automated and radiation-free diagnosis of craniosynostosis.

However, the low prevalence of craniosynostosis [12] inherent to rare diseases impedes the creation of large clinical datasets. The lack of data is further complicated by the sensitive nature of the patient recordings which show the patients' faces and are therefore subject to strict patient data restrictions. Consequently, there are no publicly available datasets of craniosynostosis patients and existing classification approaches could only be tested on in-house datasets which reduces comparability [13]. In the long run, a diagnostic prediction tool requires acceptance and trust of parents and physicians [14]. For the computer-supported diagnosis of craniosynostosis, this could be achieved by using diverse and selected datasets, by making training data accessible to independent experts, and by using classification algorithms which give an explanation or interpretation about the diagnostic prediction that they make.

Diverse approaches to overcome the described challenges have to be developed: Synthetic data might replace clinical data as it cannot be linked to an individual and can be published to increase the availability of publicly available training and evaluation data. A conversion of the 3D patient surface scans into a modality which does not disclose patient identity could facilitate data sharing. Most importantly, it is also required to develop interpretable and well-performing classification approaches suitable to use anonymous clinical data, synthetic data, and data augmentation.

1.2 State of the Art

According to bibliometric studies [15, 16], the research field of craniosynostosis is dominated by clinical, neurogenetic, and surgical publications, while engineering studies only play a minor role. CT imaging is the most established imaging modalitiy [17], while 3D surface scanning [11] has been emerging in the last decade [18].

CT imaging has been used for both assessment and computer-assisted diagnosis of craniosynostosis [17, 19]. This included studying shape differences between pathological and physiological subjects [19], statistical analysis [20], shape quantification for brain assessment of craniosynostosis patients [21], and shape analysis using atlas-based approaches [22]. Clinically recorded ratios of width, length, and diagonal measurements have been extracted automatically [23], as well as circumference-based measurement [24]. Classification approaches have been proposed for binary classification of sagittal synostosis using Fourier analysis [25], image descriptors [26], 2D skull bone projection [27], multi-view CT projection [28] with up to 90.5 % classification accuracy, and multi-class classification of craniosynostosis using a statistical shape model (SSM) and manually defined shape descriptors [29] with reported accuracies of 95.7 % on 141 cases.

The introduction of 3D surface scanners enabled the radiation-free recording of craniosynostosis patients before and after clinical intervention. In particular, SSMs were frequently used for shape quantification: Statistical differences between the principal components of a pathology-mixed model and a physiological model could

be observed [30] and combinations of principal components could be correlated with clinical length measurements [31]. SSMs and principal component analysis were also used to quantify shape changes between pre-and post-operative craniosynostosis patients using healthy shape models [32, 33] or asymmetry models [34]. Using distance-based measurements with respect to a pre-defined center point [35] is an alternative to statistical shape modeling and has been used to extract head volume [36] and to automatically compute cephalometric parameters [37]. The first classification approach using 3D surface scans [13] was published in 2020 (during the implementation of this thesis) and demonstrated a classification accuracy of 99.5 % on 196 cases using a feedforward neural network and triangular ray tracing with distance extraction. While it is one of the most promising approaches, manual alignment of the scans was required and data augmentation or synthetic data could not be added on the fly [13].

Although it has been acknowledged that a lack of data is detrimental to current classification models [38, 39], and it has been advocated for the inclusion of synthetic data [13], no study synthesized or augmented datasets using SSMs or generative adversarial networks (GANs). The creation of publicly available clinical or synthetic datasets could increase reproducibility [40] and collaboration. The lack of publicly available data also leads to in-house validation of the models, which reduces comparability, especially if no other classification method has been tested for comparison. The usage of increasingly complex neural networks (NNs) [13, 28, 38] could decrease trust of physicians and parents alike [41–43] which could be solved with interpretable models or even replacing them with inherently explainable white box models.

1.3 Research Question and Goals of This Work

The goal of this thesis is to improve data-driven and radiation-free classification of craniosynostosis by developing new methods for the classification of this condition. As clinical data are rare, the dependence on them is reduced by developing approaches for synthetic data generation and data augmentation during model training. These two problem definitions lead to the following research questions, theses and hypotheses:

Research question I

What are suitable data-driven classification methods for craniosynostosis?

Thesis I

Both neural-network-based methods and non-neural-network-based methods are capable of classifying craniosynostosis.

Hypothesis I

An F1-score of 0.95 or higher can be obtained by neural-network-based methods and non-neural-network-based methods.

Research question II

What are suitable ways to reduce the dependency on clinical data for the classification of craniosynostosis?

Thesis II

Synthetic data can replace clinical data during training with similar performance on clinical test data.

Hypothesis II

If trained on synthetic data, the F1-score of the classifier is at most 0.05 smaller compared to the classifier trained on clinical data.

Both research question are strongly rooted in practical applicability, often a typical characteristic in the field of biomedical engineering. They aim to contribute to the field of image-guided diagnostic of craniosynostosis tailored toward the clinical use-case. Both questions have the underlying assumption of a common dataset for comparison. From an engineering perspective, the following milestones are required to answer the proposed questions and to overcome the presented challenges:

- **Systematically evaluate classification approaches:**
  Develop new and improve existing classification methods and systematically compare them in terms of classification performance to answer the proposed questions.
- **Encode the 3D geometry into an anonymous representation:**
  Develop an approach to encode the 3D geometry of the head into a representation suitable for classification, e.g. a 2D image. This makes data sharing easier and might improve classification performance.

- **Synthesize pathology-specific data:**
  Translate, incorporate and develop methods to synthesize realistic and pathology-specific data (for example using an SSM). This enables the creation of synthetic datasets.
- **Increase data availability:**
  Make synthetic data publicly available, contributing to the scientific community, and enabling other scientists to test their algorithms in a controlled and comparable environment. This facilitates collaboration and makes development of applications related to craniosynostosis easier.

1.4 Structure of This Thesis

The structure, main parts and their dependencies are visually outlined in Fig. 1.1.

Part I describes the medical and mathematical fundamentals required to understand the methods and results:

- **Chapter 2** gives a brief introduction about the pathogenesis, diagnosis, and therapy of craniosynostosis.
- **Chapter 3** provides an overview about the most common ML methods for classification (NN-based and non-NN-based) as well as GAN-based data synthesis.
- **Chapter 4** introduces the concept of statistical shape modeling, including their creation and evaluation.

Part II introduces the clinical dataset to evaluate the classification algorithms, the preprocessing, and the published SSM:

- **Chapter 5** introduces the clinical dataset and preprocessing used in this thesis.
- **Chapter 6** revolves around the creation, evaluation, and publication of a pathology-specific SSM and demonstrates some of its clinical use cases.

Part III comprises a collection of proposed classification methods for craniosynostosis and contains the main methodological contributions:

- **Chapter 7** presents a multi-height classification approach based on clinically established cephalometric parameters.
- **Chapter 8** introduces a classification approach based on the shape parameter vector of a cranium SSM derived from the work in Chapter 6.
- **Chapter 9** showcases the 3D-2D conversion of the 3D head shape into a 2D image in combination with a classification approach based on a convolutional neural network (CNN).

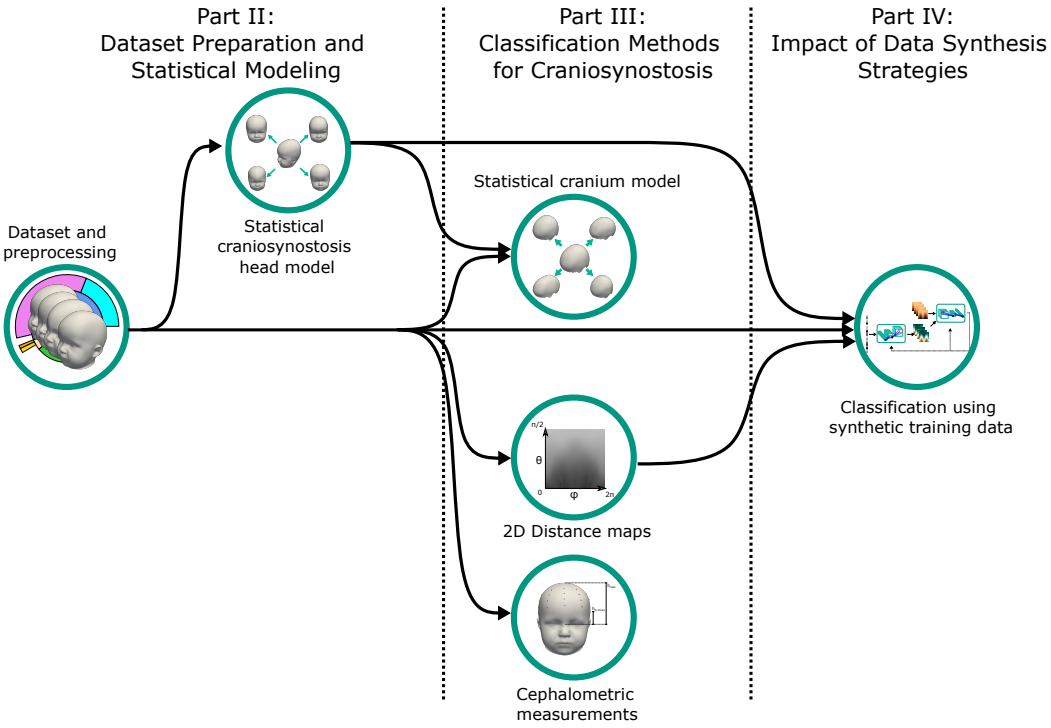


Figure 1.1: This schematic shows the main parts of this work with their dependencies and their relationship to the final experiment. From left to right: Part II, III, and IV.

Part IV combines the CNN-based classification method with multiple data synthesis approaches to replace clinical data with synthetic data:

- **Chapter 10** presents a data synthesis pipeline and performs a classification of craniosynostosis on synthetic data alone.

Part V outlines follow-up-studies, summarizes the main findings, and highlights the consequences for the field:

- **Chapter 11** proposes the immediate follow-up steps for clinical applicability and explores possible future paths.
- **Chapter 12** concludes this thesis by summarizing the contributions and answering the scientific questions posed in the introduction.

PART I

FUNDAMENTALS

Chapter 2

Medical Fundamentals

2.1 Craniosynostosis and Head Deformities

2.1.1 Skull Development in Infants

During the first years of life, the human brain grows rapidly and requires sufficient space to expand. The neurocranium is the protective case of the brain and, together with the facial skeleton, forms the human skull. Eight bones are considered to be part of the neurocranium and the four largest ones (frontal bone, the left and right parietal bones, and the occipital bone) form the calvaria, the top part of the skull [44].

The cranial sutures (depicted in Fig. 2.1) occupy the space between the cranial bones and consist of collagen fibres which allow tiny movements essential for the growth of the skull and head. After a couple of months, the sutures begin to ossify and have turned to bone after one to two years [45]. During regular growth, the lambdoid suture starts fusing after 2–3 months, the metopic suture after 3–9 months, and the sagittal and the coronal sutures after 18–24 months [46]. The ossification continues the whole lifetime and different degrees of ossification are even considered for forensic age estimation [47]. Brain growth is considered the driver for calvarial expansion, therefore it is essential for a uniform and regular head growth to allow bone movement and expansion which requires that the sutures are not yet fully ossified. If the closure of one or multiple skull sutures happens prematurely, this is called craniosynostosis.

2.1.2 Pathology and Pathogenesis of Craniosynostosis

Craniosynostosis is characterized by the premature ossification of skull sutures in infants and results in irregular growth patterns. Its reported prevalence is three to six cases per 10,000 live births [12, 48–50]. Due to the low prevalence, the American National Organization for Rare Disorders has included craniosynostosis into the list

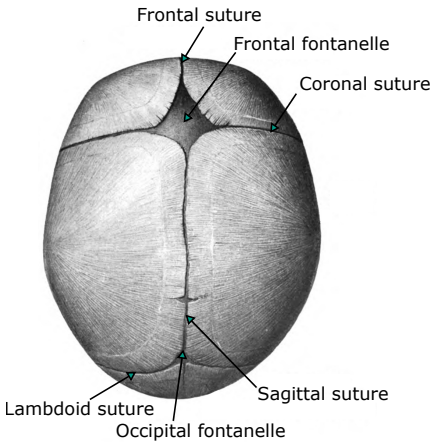


Figure 2.1: Position and names of the cranial sutures and fontanelles in a healthy infant. Adopted from [45].

of rare diseases. Head deformities related to craniosynostosis have been described since the antiquity [51]. The scientific foundation beyond simple descriptions of the resulting shape were laid by Rudolf Virchow in 1851 [52, 53], who linked the head deformities to the closure of skull sutures and created the scientific foundation of this disease. He hypothesized the closure of sutures leads to compensatory growth perpendicular to the suture, summarized as Virchow's law, which still holds up to date.

Craniosynostosis can occur isolated (affecting one suture) or non-isolated (affecting multiple sutures). Multi-suture synostosis is the minority of cases and accounts for approximately 5 to 20 % of all occurrences, the numbers vary depending on the population, region, and medical access of the community [50, 54]. The causes of multi-suture synostosis are often syndromic conditions such as Crouzon, Muenke, or Pfeiffer syndromes, which have genetic reasons and tend to show multiple and distinct craniofacial features [55]. Isolated craniosynostosis is the most common type of craniosynostosis and its causes are believed to be multifactorial: Pathogenetic research suggests primarily genetic involvement due to the increased occurrence of craniosynostosis in the same families [50]. Some genetic mutations have been identified to cause premature fusion of specific sutures [56]. Other risk factors such as smoking during pregnancy have been attributed to increased risk of craniosynostosis [57].

Isolated craniosynostosis can be classified into sagittal synostosis (scaphocephaly), metopic synostosis (trigonocephaly), unilateral coronary synostosis (anterior plagiocephaly), lambda synostosis (posterior plagiocephaly) and bicoronal synostosis (brachycephaly).[1] The most common types are sagittal synostosis with around 50 % [54] and metopic synostosis with around 30 % of all cases [50], while uni- and

[1]Although bi-coronal synostosis involves two sutures, the medical community counts it as an isolated type of craniosynostosis.

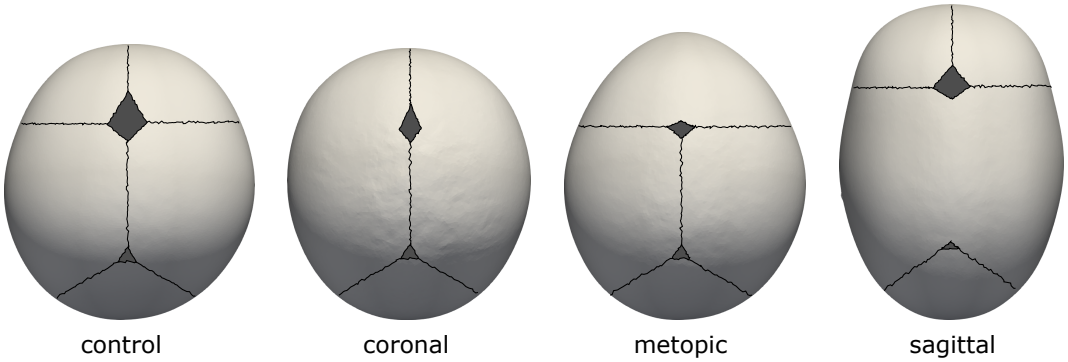


Figure 2.2: Sutures and their closure and corresponding head deformities associated with them that will be relevant to this work.

bi-coronal synostosis together account for around 15–20 % of all cases [58], and lambda synostosis around 2 % [58]. Symptoms of isolated craniosynostosis are a deformity of the neurocranium and consecutively deformations of the viscerocranium (the ears, facial and dental parts) can be observed.

Craniosynostosis can cause aesthetic, acoustic, ophthalmological complications and, most severely, has been linked to elevated intracranial pressure [5, 59], which can lead to reduced brain growth and reduced neuropsychological development such as vision impairment and slowed development of language, speech, and visual spatial skills [6, 7]. However, as the grow conditions for the brain worsen with time, the risk of a reduced mental development and impairment is the main reason why parents and physicians try to act fast. Early diagnosis and effective treatment are therefore key.

2.1.3 Diagnosis and Treatment of Craniosynostosis

As dictated by Virchow's Law, the premature closure of a suture limits the expansion of the skull perpendicular to the fused suture, causing compensatory growth along the suture, resulting in distinct head shapes [53], depicted in Fig. 2.2. In clinical practice, diagnosis is performed in specialized medical hospitals and consists of visual examination, palpation of the suture positions and fontanels, cephalometric measurements, and medical imaging.

Computed tomography (CT) imaging is the gold standard for diagnosis as well as surgical planning and is still routinely performed in many craniofacial centers worldwide. However, this exposes infants to ionizing radiation which should be avoided [8]. Plain radiography can sometimes be used to reduce the impact of radiation, but is less common [10]. Alternative imaging methods include black bone magnetic resonance imaging (MRI) [60], ultrasound sonographic imaging [61], and 3D photography. MRI has the notable drawback that the infant needs to be sedated during image acquisition to prevent moving [60, 62]. Ultrasound imaging and 3D

photography are radiation-free and broadly available diagnostic options. Ultrasound sonographic imaging is the most operator-dependent imaging routine and requires a well-trained expert to visualize the entire suture and to correctly interpret the image [10]. As a radiation-free alternative to traditional CT, 3D surface scans provide inexpensive and fast means to objectively quantify head shape without exposure to radiation or sedation. 3D surface scan recordings are typically used to monitor the condition before surgery and the head development after the operation [11].

Surgical treatment involves resection of the synostosis as well as remodeling and reshaping of the cranial vault. The operation aims to prevent abnormal brain growth, thus enabling a regular development of skull and face [8, 63]. The operation is often performed during the first year of life to ensure sufficient re-ossification and to reduce strong deformations in the first place [9]. The introduction of several surgical methods has made the surgical treatment safe since the 1960s [51]. Complications during surgery are rare [64] and in most cases a normalized head shape is achieved [65]. For further reading, it is referred to [66].

2.1.4 Non-synostotic Positional Plagiocephaly

The most important differential diagnosis for craniosynostosis are head deformities without suture fusion. These head deformities are mainly manifested as a non-synostotic posterior plagiocephaly. As opposed to craniosynostosis, causes of positional plagiocephaly are mostly caused by environmental circumstances such as unilateral pressure such as static positioning. Some children have the tendency to lie on one side which can contribute to positional plagiocephaly [67]. The success of the "back to sleep" campaign in the United States to prevent sudden infant death syndrome (SIDS) is widely attributed to a rise of positional plagiocephaly cases in infants [6, 50, 67, 68]. Back sleeping is one of the most effective ways to prevent SIDS which inadvertently also led to an increased incidence of positional plagiocephaly. Positional skull deformities are generally benign, reversible, and do not require surgical intervention. They are often treated with positioning pillows, helmet therapy or changes in positioning behavior [69, 70].

Positional plagiocephaly should not be confused with coronary synostosis or lambda synostosis. As positional plagiocephaly is more common than both coronary and lambda synostosis combined, the term "plagiocephaly" often refers to positional plagiocephaly.

2.2 Cephalometric Measurements

The two most important metrics for the evaluation of head asymmetry and shape are cephalic index (CI) (also sometimes referred to as "cephalic ratio") and cranial

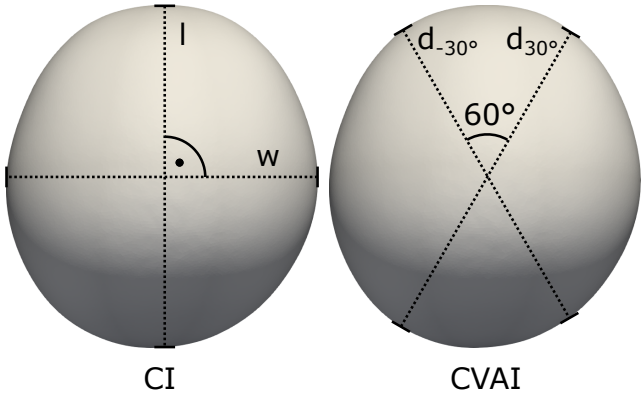


Figure 2.3: Measurement visualization of width w, length l, and the two diagonals d_{30° and d_{-30° to compute cephalic index (CI) and cranial vault asymmetry index (CVAI) on the head viewed from top.

vault asymmetry index (CVAI). Both measurements are depicted in Fig. 2.3. CI describes the ratio of width w over length l of the head, computed as

$$\mathrm{CI} = \frac{w}{l}. \tag{2.1}$$

In the medical community, the measurement is often multiplied by 100 % to avoid a range close to 1. The CI is influenced by several craniofacial deformities such as sagittal synostosis and coronal synostosis. CVAI is a measure of asymmetry and is computed at ±30° [71] on the diagonals at d_{30° and d_{-30° as

$$\mathrm{CVAI} = \frac{d_{-30^\circ} - d_{30^\circ}}{\max(d_{-30^\circ}, d_{30^\circ})}. \tag{2.2}$$

CVAI is influenced especially by plagiocephaly and lambdoid synostosis, but also to a lesser extend by metopic synostosis. A visualization of the measured diagonals is provided in Fig. 2.3.

The physiological range of the mean CI for infants has been reported to be roughly between 0.79 and 0.92 and depends on age, extraction position, and ethnicity [23, 72–78]. Early measurements used calipers, but CT and 3D surface scan imaging enable computer-assisted determination of cephalometric measurements.

Chapter 3

Machine Learning Fundamentals

3.1 Non-Neural-Network-Based Machine Learning Models

This section gives a brief overview about popular supervised machine learning (ML) models such as kernel-based, probabilistic, and tree-based models which are sometimes called "traditional" ML models [79, 80]. Neural network (NN) classification models are described in Section 3.2. First ML studies focused on learning games such as chess or checkers [81] for demonstration purposes and have since been applied to many different tasks such as classification, regression, pattern recognition, and feature extraction. ML describes the approach of a machine to "learn" a representation of data or task instead of explicitly defining the parameters of the machine learning model. This avoids time-expensive manual parameter tuning for similar models operating on different domains. Instead, only a training algorithm is provided allowing the model to be optimized automatically.

Supervised learning describes the strategy to use input data with associated ground truth labels on which the model can be evaluated and optimized. This requires only a nominal scale of the input data (i.e., the labels represent categories) and is often performed for classification tasks (e.g., predicting the type of craniofacial deformity given an input sample). Many ML models are available in free and open-source software (FOSS) libraries such as `scikit-learn`. Those libraries are well documented and enable using powerful ML models for novice and experienced users.

3.1.1 Decision Trees and Random Forests

Decision trees (DTs) [82] use a hierarchical, tree-based structure to infer decision rules on the input data for classification. They are white box-classifiers and can be visualized in a tree-like if-else structure which can easily be explained or interpreted.

DTs operate directly on the input features and can easily over-fit, especially if a high tree depth is used which leads to many fine-grained distinctions. For stabilization, they are often used after feature extraction methods such as principal component analysis (PCA).

Bagged (**b**ootstrap **agg**regated) models are an ensemble of multiple decision tree models with randomly sampled input parameters which use a voting procedure (e.g., majority voting or average voting) to determine the final model output. This introduces additional randomness into the model making it less prone to over-fitting. Bagged tree models are also called random forests (RFs) [83].[1]

3.1.2 k-Nearest Neighbors

The principle behind k-nearest-neighbors (kNN) classification [84] is to classify samples based on the similarity to all training data. The sample is placed into the majority class of the k nearest neighbors. The distance metric to determine the neighbors can vary, but Euclidean distance is a common choice. kNN classification does not make assumptions with respect to the data distribution and simply contains a lookup-table of the training data and computes the distances in each query. kNN classifiers are therefore fully explainable and the closest samples can be retrieved for comparison purposes. kNN classification can also be used in combination with dimensionality reduction approaches such as PCA.

3.1.3 Naïve Bayes

Naïve Bayes (NB) classifiers use Bayes' theorem to classify samples with the "naïve" assumption that there is conditional independence of feature pairs (which is almost never the case for real-world problems). This corresponds to modeling the covariance matrix as a diagonal matrix. As a consequence regarding the assumption of conditionally independent feature pairs, maximum a posteriori (MAP) estimation can be used to classify new samples. The underlying probability function is usually modeled as a Gaussian distribution, but other distributions are also possible. Due to its simplicity, NB is a fast and robust estimator. Some considerations on why NB works well is described in [85, 86].

3.1.4 Linear Discriminant Analysis

Linear discriminant analysis (LDA) [87] uses decision boundaries to classify data according to the maximum posterior probability in a similar fashion to NB. In contrast to NB, the input features are not necessarily assumed to be conditionally

[1] Since a forest is made out of multiple trees (in this case, decision trees)

independent (which leads to a non-diagonal sample covariance matrix). LDA assumes that the different classes share the same covariance matrix which leads to a linear decision boundary between classes (hence the name). If different covariance matrices are assumed for each class, the classification decision boundary becomes quadratic resulting in quadratic discriminant analysis (QDA).

3.1.5 Support Vector Machines

Support vector machines (SVMs) [88] are classifiers that use kernel functions to transform the input features into a high-dimensional space in which the different classes can be separated by hyper-planes. They use a subset of training samples (support vectors) to construct the separating hyper-planes for the decision function and are sensitive to differently scaled inputs.

SVMs alone are inherently binary and can only construct a decision boundary between two classes. To adapt them for multi-class-problems, either a one-vs-one scheme with $n \cdot (n-1)/2$ SVMs or a one-vs-rest scheme with n SVMs has to be implemented.

3.1.6 Visualization

Prediction functions of the introduced classifiers are shown in Fig. 3.1 with different hyper-parameters for a more intuitive understanding of their behavior. The if-else decisions of the DT leads to coarse horizontal and vertical boundaries for the DT, which is very fine-grained for the RF. The prediction boundary for kNN becomes smoother for a larger number of neighbors. For LDA, the linear decision boundary is replaced by a quadratic boundary if the covariance matrices are allowed to differ. For the SVM, the kernel choice has a strong influence of on the resulting predictions.

3.2 Feedforward Neural Networks

Artificial NNs have become a popular and powerful machine learning tool for classification and regression. They are mostly inspired by the neural networks of the nervous system of animals and describe a hierarchical structure of layers which "learn" features with increasing complexity. The trainable parameters of the networks consist of weights and biases of the connections between each of the neurons. Feedforward neural networks (FNNs) are one of the fundamental types of artificial NN and have been adapted to more specialized types of networks such as convolutional neural networks (CNNs). There is excellent literature about ML with different levels of depth available (e.g. [79, 80]).

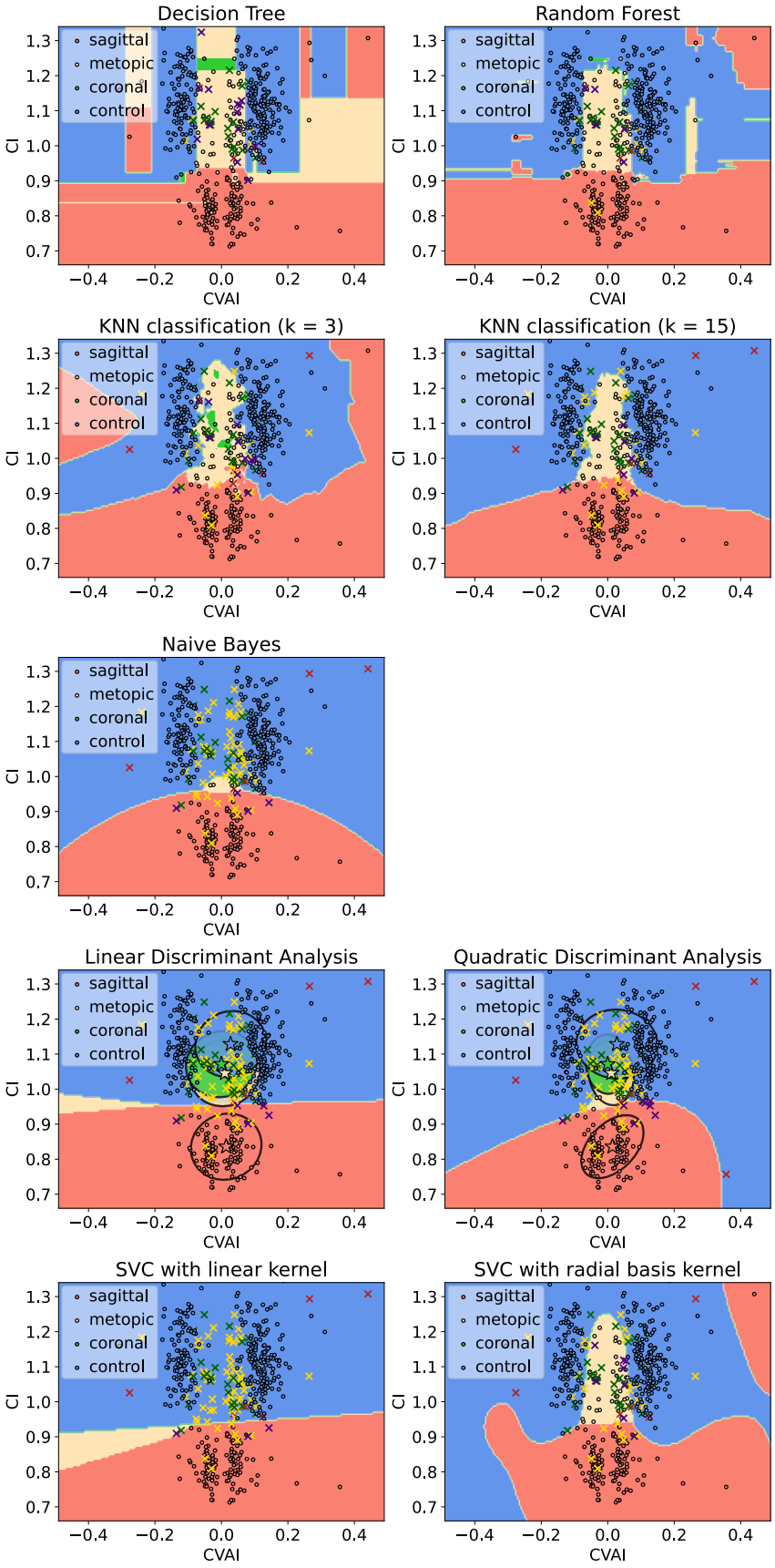


Figure 3.1: Exemplifying plot of the different classifiers to visualize their behavior. Shown is a visualization of the prediction function of a multi-class classification of the mentioned classifiers for visualization purposes only. For LDA, the covariance matrices are indicated with the ellipse and the mean with a star.

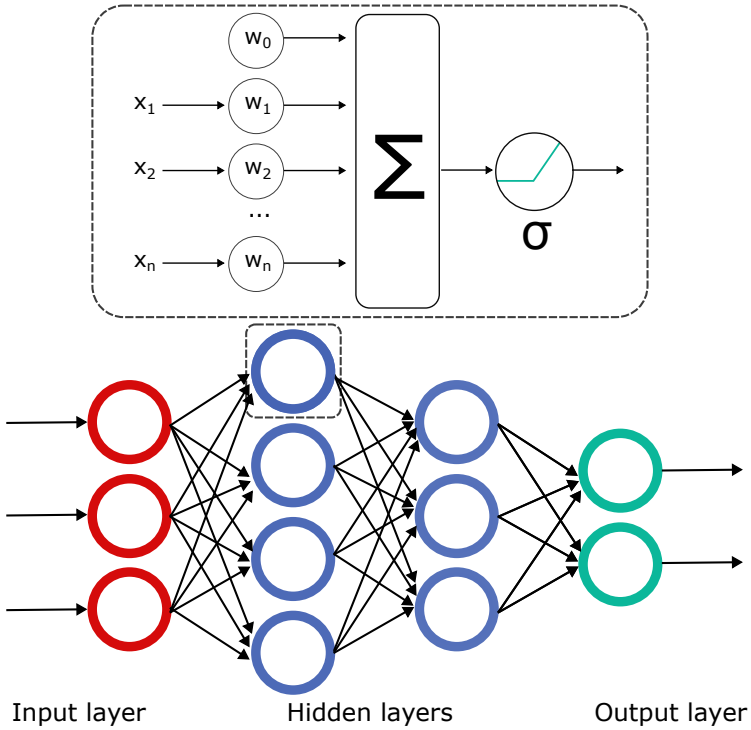


Figure 3.2: The basic structure of a neural network. A single artificial neuron receives inputs from the previous layer (above) and computes its activation value according to Eq. (3.1). This is performed for all neurons in all layers of the feedforward neural network (FNN) until the input is propagated to the output (below).

3.2.1 Structure

The fundamental building blocks of FNNs are artificial neurons, visualized in Fig. 3.2. They receive weighted inputs and pass their sum into a nonlinear activation function, mathematically described as

$$f(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{i=0}^{n} x_i \cdot w_i\right), \tag{3.1}$$

with w_i denoting the trainable network weights, x_i the respective inputs from previous layers (with $x_0 = 1$, making w_0 a trainable constant initial activation or "bias" for each neuron), and σ the activation function, for example rectified linear unit (ReLU) with $\sigma(x) = \max(0, x)$. The non-linearity of the activation function is the key element to fit the model to any type of input data.[2]

[2]In contrast, a linear activation function would lead to a system of only linear blocks which — according to the mathematical definition of linearity — would also be linear and (regardless of the number of neurons and layers) could be mathematically simplified and collapsed into one single linear layer.

Multiple neurons form a layer in which each neuron receives inputs from all previous neurons from the previous layer (see Fig. 3.2). The intermediate layers not connected to the input or output are called hidden layers. A network consisting of multiple fully connected layers is called a FNN or multi layer perceptron (MLP), making MLPs one of the simplest types of NN. As NNs can have thousands or even millions of tunable model parameters, the effective and efficient optimization of those parameters can be a substantial challenge. The model optimization process of NNs is usually referred to as model training.

3.2.2 Model Training

The objective during model training is to find the best value of the objective function, i.e., the global minimum of the cost function. Unlike some of the "traditional" ML approaches such as LDA, FNNs do not have a convex loss function, so multiple local minima exist and the random network initialization and optimization strategy affect the result. The cost function has a parameter space with a large number of parameters for optimization and the training goal is to find the parameters which yield the smallest cost function value. Such optimization can be performed by computing the gradient on the current position and descending along the negative gradient in an interval dictated by the learning rate. Since the objective function depends on all training samples, the exact gradient is costly to compute and is therefore approximated by taking only a subset of the training instances (minibatch). This training approach is called stochastic gradient descent (SGD) and is the basis for many other optimization algorithms such as adaptive moment estimation (ADAM). Often, regularization such as Tikhonov regularization (or $\mathbf{L_2}$ regularization) is performed. In the context of machine learning, this is denoted as "weight decay" since it results in a multiplicative factor to shrink the learning rate during each iteration [79].[3] During supervised training, labeled training images are presented to the mathematical model and their loss is computed to adjust the model parameters according to the gradient of the loss function with respect to the given parameter. The loss and gradients are therefore propagated backwards (hence the name back-propagation) to the first layer adjusting all available parameters according to the optimizer.[4] In deep learning libraries such as `pytorch` [89], this is performed automatically during training. For example, `pytorch` uses directed acyclic graphs

[3] Tikhonov regularization adds an additional term to the cost function which relates to the model parameters, therefore preferring solutions with smaller norms. During SGD, this is equivalent to multiplicatively shrinking the weight factor by a constant during each training iteration.

[4] From the mathematical point of view, the neurons of each layer (see also Eq. 3.1) are a function composition of the functions of previous layers, so their gradients can be obtained using the chain rule.

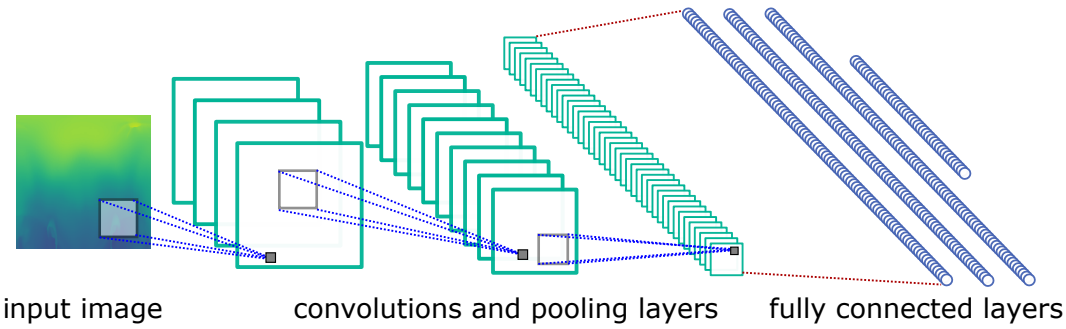


Figure 3.3: Basic structure of a convolutional neural network suitable for a classification task. Typically, convolutional layers and pooling layers are applied to the image structures until in the end, a few fully connected layers are applied for the classification.

to keep track of all values and operations used during the forward pass and employ a network graph traversal approach during back-propagation.

Unlike most non-neural-network-based classifiers, FNNs can be extensively tuned by changing hyperparameters such as number and type of model parameters (network structure, number of layers, activation function), loss function to minimize (e.g., mean squared error or cross entropy loss), optimizer (e.g., SGD or ADAM), and training (e.g., number of training iterations (epochs), number of parallel input (minibatches)).

Due to the high popularity of NNs, their versatility in many application, the vast number of different network architectures and countless possibilities of hyper-parameter-tuning, there is a vivid community and a massive amount of available information about their training and optimization.

3.3 Convolutional Neural Networks

3.3.1 Structure

CNN are widely used for image processing, segmentation, 2D classification, and computer vision. In contrast to FNNs, CNNs use convolutional filter kernels which are slid across the image. This spatially confines features in the kernels and reduces the number of trainable parameters (thus implicitly applying regularization).

Convolutional layers employ multiple filter kernels in parallel and are frequently followed by a pooling layer, which partitions the image into rectangles and yields one output per partition. The most common approach is max pooling, which passes only the largest value from its inputs. Pooling serves as a nonlinear image downsampling to extract the most important features and their relative locations. After several convolutional layers and pooling layers, usually a couple of fully connected layers are used to perform a classification on the extracted features in

which the number of neurons in the last fully connected layer correspond to the number of classes. Fig. 3.3 shows the structure of a typical CNN.

In image classification, CNNs have been extremely successful and have supplanted virtually all competing approaches [90]. Additionally, easy access through FOSS libraries and a large selection of many different pre-trained networks give incentives to change, process, or re-arrange input data to 2D images to enable the usage of CNNs in multiple domains. Some examples include classification on electrocardiographic images, from time-sensitive wavelet transformation [91], optical flow on video data for gait recognition [92], or images assembled from electrocardiographic imaging [93].

The output of the convolutional layers of the CNN often correspond to successively more complex features [94]. Empirically, those features in the first layers tend to be similar across different domains [95]. The concept of *transfer learning* exploits this concepts and uses pre-trained networks which are fine-tuned, i.e., their last layer is replaced with a layer matching the number of classes and re-trained on the current classification problem. This reduces training time and enables the creation of convolutional features even if little training data are available. There is excellent literature available for further reading [96].

3.3.2 Training Strategies

3.3.2.1 Data Subdivision

For predictive models which perform data-driven learning, a popular strategy for model evaluation is dividing the available dataset into training, validation, and test sets (e.g. 60 % — 20 % — 20 % is a typical split). In this scenario, the model is always trained on the training set, model performance is evaluated and optimized on the unseen validation set. The final evaluation is performed on the test set to avoid over-optimization toward the validation set, which ideally has never been used by the developers before.[5]

Ideally, the training, validation, and test sets have similar class distributions on each set, which can be achieved using stratification. During stratification, splitting is performed for each class individually and combined for each split (see also Fig. 3.4 for stratified cross-validation).

One disadvantage of the standard train, validation, and test split is that by chance, the model might have a "lucky" or "unlucky" test set, leading to over- or under-estimation of the model performance, especially for small datasets. Cross-validation is an option to circumvent this problem.

[5] During contests, e.g., during the Netflix price [97], the test set is usually kept secret and only used for the final scoring of all participants.

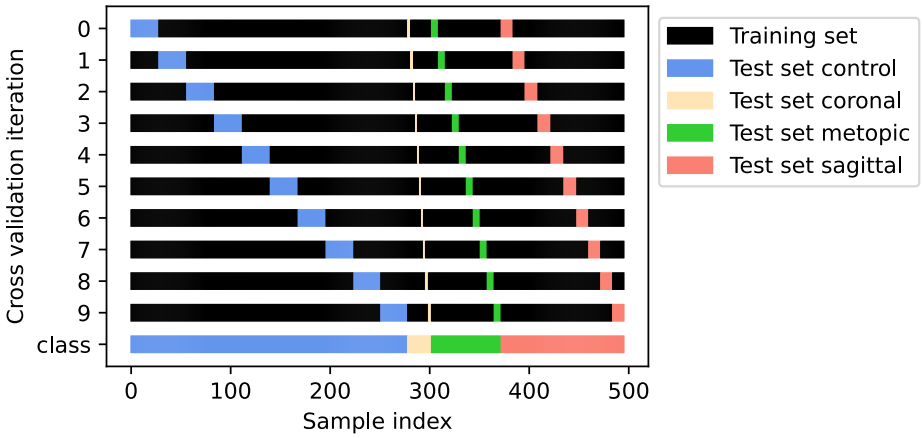


Figure 3.4: Stratified 10-fold cross-validation with a class distribution equivalent to the main dataset used in this thesis.

3.3.2.2 Stratified k-fold Cross Validation

Cross validation consists of subdividing the full dataset into a set of k splits. During k-fold cross validation, the model is tested on the kth split and trained on the remaining data. Additionally, stratification can be used to ensure the same class distribution among splits. k-fold cross validation leads to an increased training time (since k models must be trained), but ensures that each sample has been tested in one of the folds. The test set distributions of stratified 10-fold cross validation are visually exemplified in Fig. 3.4. Another advantage of cross-validation is that the performance is evaluated multiple times and therefore statistics such as mean performance and its standard deviation can be obtained giving an approximation of the robustness of the classifier.[6]

3.4 Evaluation of Classification Models

Correctly and incorrectly classified samples can be arranged in a confusion matrix according to their true and predicted label and give the full information of the classifiers. To evaluate classification performance quantitatively, several metrics exist, with some being tailored toward specific problems (e.g., class imbalance) and use-cases (e.g., information retrieval of search queries or diagnostic tests). An exemplary confusion matrix and the definitions of sensitivity, precision, and recall derived from it are displayed in Fig. 3.5.

[6]Accumulating all predictions and computing a metric from the combined predictions instead of computing mean and standard deviation is usually not a valid way to measure classification performance since it often over-estimates performance on non-linear metrics.

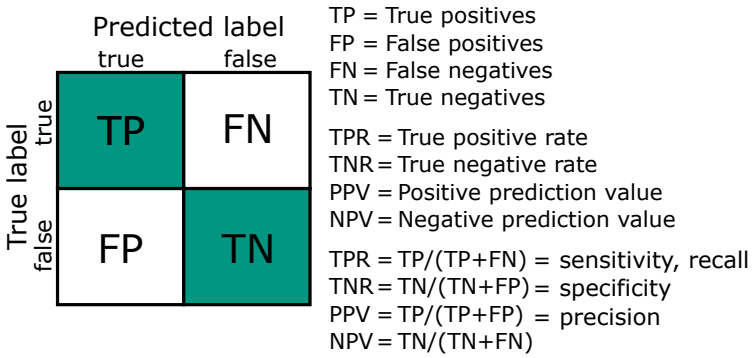


Figure 3.5: Confusion matrix definition including true positives, true negatives, false positives, and false negatives. The definitions of true positive rate, true negative rate, positive predictive value, negative predictive value are included.

Accuracy is computed as the ratio of correctly classified instances over all instances. F1-score can be calculated as the harmonic mean of precision and recall and can therefore also be computed when the number of true and false negatives is unknown (e.g., during data retrieval). G-mean describes the geometric mean of all class-wise sensitivities (which, in a binary classification problem is equal to sensitivity and specificity). In a binary classification problem, the three metrics can be computed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{3.2}$$

$$\text{F1-score} = 2 \cdot \frac{\text{TPR} \cdot \text{PPV}}{\text{TPR} + \text{PPV}}, \tag{3.3}$$

$$\text{G-mean} = \sqrt{\text{TPR} \cdot \text{TNR}}. \tag{3.4}$$

$$\tag{3.5}$$

While accuracy is likely the most common and intuitive metric of the three, it is severely influenced by the class distribution of the dataset and given an imbalanced dataset: high accuracy values can be achieved by classifiying everything as the majority class and do not necessarily represent a robust classifier. The F1-score is regarded a more suitable metric when dealing with imbalanced datasets, because class imbalance influences it to a lesser extend.[7] G-mean is independent of class distribution [98] and its value cannot be higher than the lowest sensitivity. However, its non-linearity and high sensitivity to minority classes makes it less intuitive and is therefore less popular. For non-binary classification problems, the F1-score of each class can be computed as the arithmetic mean of individual binary F1-scores (macro F1-score).

[7]However, the F1-score has other shortcomings: As the F1-score relies on precision, it is not symmetric and therefore depends on the definition of which is the "positive" class. In other words, the F1-score can be manipulated for binary classification problems by swapping the positive and negative classes. However, this behavior disappears for a multi-class problem.

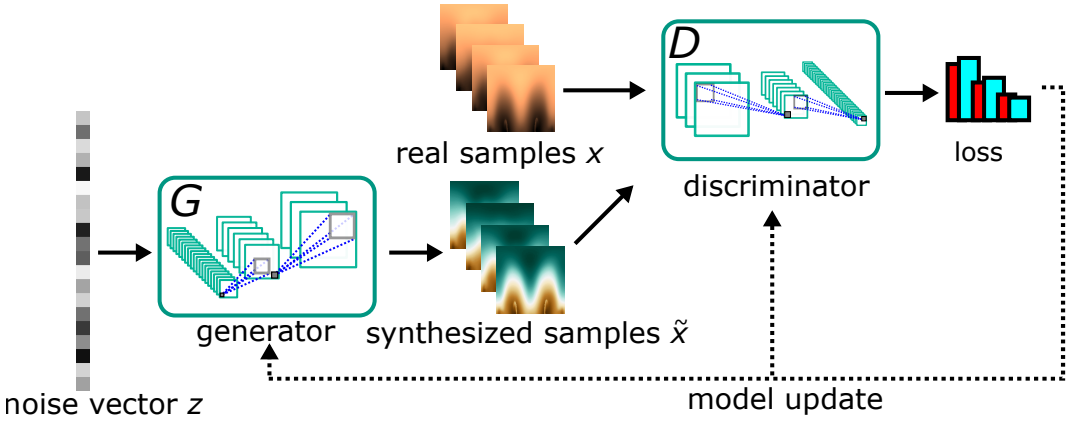


Figure 3.6: GANs training visualized: The generator constructs images from random noise, while the discriminator decides if the images are real or fake. This creates a loss which can be used to update both networks.

3.5 Generative Adversarial Networks

Generative NNs such as generative adversarial networks (GANs) [99] are designed to model the data distribution $p(x)$ of a given dataset to be able to synthesize data from the approximated distribution.[8]

GANs have two components, a generator and a discriminator, which are trained simultaneously and compete against each other in a zero-sum game. The generator network receives random noise z and synthesizes random images $\tilde{x}(z)$, while the discriminator network receives the generator's images and the real samples trying to predict which images are real and which are forged. Successively, the generator synthesizes images more and more similar to the sample distribution $p(x)$, and the discriminator becomes better at distinguishing between them. From a game-theoretic standpoint, they optimize their strategy to minimize their individual loss until a deviation from this strategy would lead to a higher individual loss and a Nash equilibrium is reached. Model training is visualized in Fig. 3.6. According to [99], the training equation can be described as

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_\mathrm{r}} \log D(x) + \mathbb{E}_{z\sim p_z} \log(1 - D(G(z))). \tag{3.6}$$

x denotes the real data and $D(x)$ the probability that x came from the real data distribution p_r, while $G(z)$ denotes the generator creating an image from a noise vector z sampled from the noise distribution p_z. $\mathbb{E}$ denotes expected values of the data distribution. The training of GANs is considered challenging since the two adversaries have to be trained simultaneously. Typical problems which can arise

[8] During classification problems, the conditional probability function $p(x|y)$ is modeled (predict x given an input y).

are vanishing gradients (which effectively stops training for the generator) or mode collapse (in which only one subset of classes is generated).

As with deep learning, the tuning possibilities of GANs are immense and specialized approaches for specific problems exist, mostly for, but not limited to, image generation and image modification. Examples include image synthesis using deep convolutional generative adversarial networks (DCGANs) [100], domain transfer such as style transfer [101–103], and image context encoding for inpainting [104]. Two variations of GANs are described below.

Conditional GAN The notion of conditional GANs (cGANs) [105] is to add a conditional component to the inputs of both generator and discriminator so that they can be conditioned to behave differently according to an input label y. In practice, this is usually implemented as an embedding vector which serves as an additional input parameter fed into both the generator and the discriminator. This changes the cost function (compared to Eq. 3.6):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_\mathrm{r}} \log D(x|y) + \mathbb{E}_{z\sim p_z} \log(1 - D(G(z|y))) \tag{3.7}$$

Wasserstein GAN Wasserstein generative adversarial networks (WGANs) [106] aim to provide a continuous gradient for a larger group of distributions to avoid vanishing gradients. While they originally proposed weight clipping, a regularization term controlled by a parameter λ can also be used [107], yielding the loss L_W as

$$L_W(D,G) = \mathbb{E}_{\tilde{x}\sim p_z} D(\tilde{x}) - \mathbb{E}_{x\sim p_\mathrm{r}} D(x) + \lambda \mathbb{E}_{\hat{x}\sim p_{\hat{x}}} (||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2, \tag{3.8}$$

with x denoting the real samples, $\tilde{x}$ denoting the generator samples from $G(z)$, and $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$ with ϵ denoting a uniformly distributed random variable between 0 and 1. The penalty term samples from a distribution $\hat{x}$ between the generator and discriminator, and, according to the authors, experimentally resulted in good performance [107].

Chapter 4

Statistical Shape Modeling Fundamentals

4.1 Model Construction

A statistical shape model (SSM) describes the shape variability of a geometrical object by means of mathematical deformations in terms of probability and statistics. Its applications range from shape reconstruction using partially available data [108] over shape registration [109] to shape analysis [30] and data synthesis [110]. Mathematical definitions of shape [111] define it as the characteristic form that remains when position, orientation, reflection, and scale are removed. In the medical context, this definition can be more lenient and often includes scale because it is related to age and is associated with different features (e.g., when considering an infant head and an elderly head).

Point distribution models (PDMs) describe the statistical information incorporated in the model by a set of points (often in a triangular mesh) and are the most common type of SSMs. Cootes et al. [112, 113] is credited as pioneering the creation of the first SSM using training data (the available data from which the statistical information is derived). Human face and head models [33, 110, 114] are one of the benchmarks applications due to their large variety in applications and the (admittedly less quantifiable) human fascination of modeling the human face due to its meaningfulness in human interaction, recognition, and emotion. Popular models of the human face or head available upon request include the Basel face model [110] and the Liverpool-York-head-model [33, 115]. Combinations of SSMs unify separate models to increase statistical variety [116, 117]. For an overview about the current trends and challenges, the reader is referred to [118].

Figure 4.1 visualizes the required steps for model construction and shows the main steps required to construct a PDM. During this thesis, each dataset sample $\mathbf{\Gamma}_i \in \mathbb{R}^{p_i \times 3}$ is a triangular 3D surface mesh, each having a different number of points

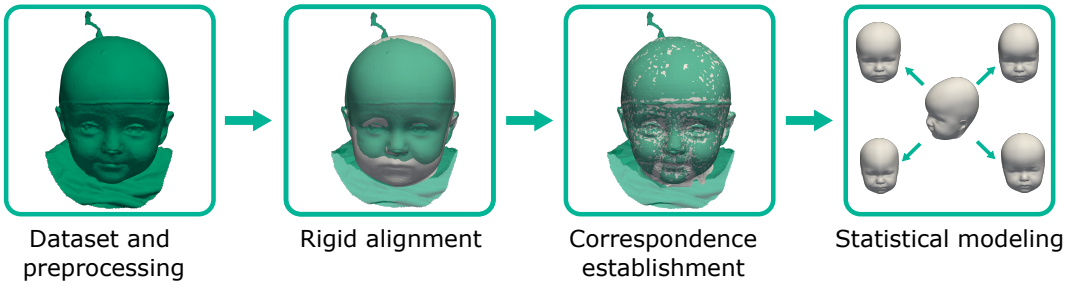


Figure 4.1: The statistical shape model pipeline employed in this study. The target scan is colored green with the deforming template in white.

p_i and cells.[1] As a prerequisite to derive statistical information from the training data, it is required that the subjects share the same point identifiers across all scans so that each identifier has a unique morphological meaning. In practice, this requires the use of a reference shape or template $\mathbf{\Gamma}_\mathrm{t} \in \mathbb{R}^{3\times p_\mathrm{r}}$ which is deformed to match each sampe and usually has a different number of points p_r. The template can be any shape, but is often either one shape of the training shapes or a related and similar shape (e.g., a smoothed version with symmetric vertices). The following steps need to be performed on each dataset sample on the template.

- Dataset acquisition and preprocessing of all scans $\mathbf{\Gamma}_i$, which includes removing artifacts such as duplicate vertices, etc.
- Rigid alignment to obtain the aligned shape $\mathbf{\Gamma}_i^\mathrm{a}$ to match the same morphological regions of template and sample. Procrustes analysis (see Section 4.2.1) is one of the typical tools for this task. Landmark-free registration algorithms such as random sample consensus (RANSAC) [119] can also be used, but are less robust.
- Correspondences establishment (see Section 4.3) is the process to obtain a common representation $\mathbf{\Gamma}_i^\mathrm{c}$, suitable for statistical analysis. In contrast to the alignment step before, this involves nonrigid shape deformations and can be performed using shape morphing.[2]
- Rigid alignment is performed using 4.2.2 to remove rotational and translational elements introduced during establishing correspondence.
- Statistical analysis is performed last and consists of principal component analysis (PCA) [121, 122]. This enables a representation in which model instances can be synthesized and matched in terms of a normally distributed shape vector.

[1] Cells (or faces) are usually not modeled because they do not provide shape information.
[2] Some methods such as nonrigid iterative closest points affine (ICPA) [120] can also perform alignment and correspondence establishment simultaneously.

4.2 Spatial Alignment

In a first step, template and target have to be aligned. If landmarks are available, it is often preferred to make use of them using Procrustes analysis (see 4.2.1). Facial landmark detectors [123, 124] can provide a set of landmarks if required. This step removes rotation and translation and is crucial for correspondence retrieval which relies on close morphologically equal regions.

4.2.1 Ordinary and Orthogonal Procrustes Analysis

Procrustes analysis is an algorithm to align two shapes with a set of known correspondences subject to a distance metric such as the Euclidean distance. Depending on the use case, Procrustes analysis can make use of rotation, translation, reflection, and uniform scaling to best fit the moving set $\mathbf{X}_\mathrm{m} \in \mathbb{R}^{p\times 3}$ to the fixed set $\mathbf{X}_\mathrm{f} \in \mathbb{R}^{p\times 3}$.

The translational component can be removed by subtracting the centroids (computed as the arithmetic mean), aligning both shapes to the origin and re-scaling them using their Frobenius norm:

$$\mathbf{X}_\mathrm{m}^\mathrm{t} = \mathbf{X}_\mathrm{m} - \overline{\mathbf{X}}_\mathrm{m} \tag{4.1}$$

$$\mathbf{X}_\mathrm{f}^\mathrm{t} = \mathbf{X}_\mathrm{f} - \overline{\mathbf{X}}_\mathrm{f} \tag{4.2}$$

$$\mathbf{X}_\mathrm{m}^\mathrm{s} = \mathbf{X}_\mathrm{m}^\mathrm{t}/\mathrm{s}_\mathrm{m} \quad \text{with } \mathrm{s}_\mathrm{m} = ||\mathbf{X}_\mathrm{m}^\mathrm{t}||_F \tag{4.3}$$

$$\mathbf{X}_\mathrm{f}^\mathrm{s} = \mathbf{X}_\mathrm{f}^\mathrm{t}/\mathrm{s}_\mathrm{f} \quad \text{with } \mathrm{s}_\mathrm{f} = ||\mathbf{X}_\mathrm{f}^\mathrm{t}||_F \tag{4.4}$$

The rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ can be computed efficiently by exploiting the singular value decomposition (SVD):

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} = {\mathbf{X}_\mathrm{m}^\mathrm{s}}^\mathsf{T} \cdot \mathbf{X}_\mathrm{f}^\mathrm{s} \tag{4.5}$$

$$\mathbf{R} = \mathbf{V} \cdot \mathbf{U}^\mathsf{T} \tag{4.6}$$

The remaining scaling is contained in $\mathbf{\Sigma} \in \mathbb{R}^{3\times 3}$. The transformation of the moving points to the target points is as follows:

$$\mathbf{X}_\mathrm{m}^\mathrm{a} = \mathbf{X}_\mathrm{m} \cdot \mathbf{R} \cdot \frac{s_\mathrm{f}}{s_\mathrm{m}} \cdot \mathrm{trace}(\mathbf{\Sigma}) + \overline{\mathbf{X}}_\mathrm{f} - \mathbf{X}_\mathrm{m}^\mathrm{t} \cdot \mathbf{R} \tag{4.7}$$

If only rotation and translation are removed, the approach is called the *orthogonal* Procrustes analysis (as the resulting transformation will be orthogonal). No scaling is performed in orthogonal Procrustes Analysis. If no reflection is desired, the determinant of the matrix has to be $\det(\mathbf{R}) = 1$ (when reflecting, $\det(\mathbf{R}) = -1$).

Table 4.1: Pseudocode of GPA to align all morphed subjects to each other using orthogonal Procrustes analysis.

1:	Choose arbitrary reference shape $\mathbf{\Gamma}^{\mathrm{r}}$ from the set
2:	Until convergence of the reference shape, do:
3:	Align all shapes: $\mathbf{\Gamma}_{\mathrm{i}}^{\mathrm{a}} = \mathrm{OrthogonalProcrustes}(\mathbf{\Gamma}_{\mathrm{i}}, \mathbf{\Gamma}^{\mathrm{r}})$
4:	Compute mean aligned shape $\overline{\mathbf{\Gamma}}^{\mathrm{a}}$
5:	Set the mean shape as the new reference shape: $\mathbf{\Gamma}^{\mathrm{r}} = \overline{\mathbf{\Gamma}}^{\mathrm{a}}$

4.2.2 Generalized Procrustes Analysis

In contrast to ordinary Procrustes analysis, which registers one shape to a reference shape, generalized Procrustes analysis (GPA) registers a set of shapes without a predefined reference. GPA performs Procrustes analysis iteratively and aligns all shapes to the mean shape of the set. Multiple options for implementation exist [125], but one basic structure is described in Tab. 4.1.

4.3 Correspondence Establishment

4.3.1 Overview

"Correspondence" refers to using the same point identifiers across all scans which enables statistical analysis. After the initial rigid alignment, the template has to be mapped to each of the target scans to obtain the correspondences. This is a nonrigid surface registration problem and a variety of algorithms to solve this problem exist [126]. Many of these algorithms are tailored toward specific use-cases. Some popular approaches for correspondence establishment in the shape model community are nonrigid coherent point drift (CPD) [127], iterative closest points (ICP) variations [128, 129], and Gaussian process morphable models [109]. For head modeling, some of the most common algorithms include Laplace-Beltrami regularized projection (LBRP) [115] (see in Section 4.3.2), iterative coherent point drift (ICPD) [33], the open framework [114], including ICPA [120], nonrigid iterative closest point translation (ICPT) [120, 130]. During correspondence retrieval, the reference or template shape usually gets deformed in such a way that it "best" matches the target shape according to a cost function which depends on the used algorithm. This often introduces small translations, rotations, and locally anisotropic deformations in each iteration to the reference shape. The deformed and mapped reference shape either replaces the target shape for the further pipeline or the nearest neighbors or closest points to the surface are mapped for this purpose. The resulting shape in so-called dense correspondence $\mathbf{\Gamma}_i^{\mathrm{c}}$ has the same point identifiers as the reference shape but is in the shape of the target shape $\mathbf{\Gamma}_i^{\mathrm{c}} \in \mathbb{R}^{3\times p_{\mathrm{r}}}$.

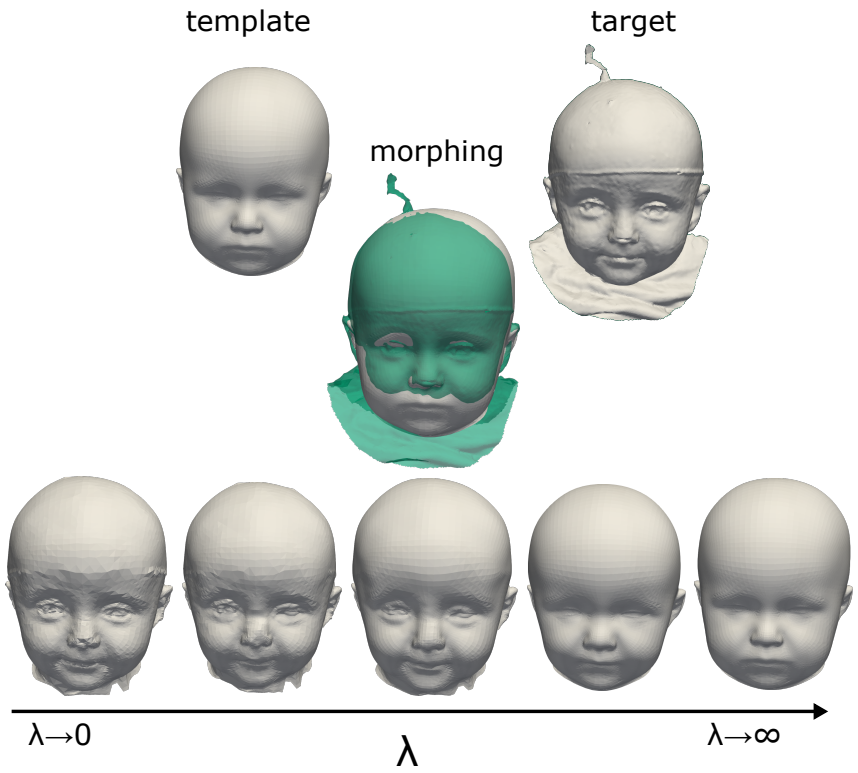


Figure 4.2: Variation of the stiffness parameter λ during template morphing.

4.3.2 Laplace-Beltrami Shape Morphing

Laplace-Beltrami regularized projection (LBRP) [33, 115] for shape morphing relies on mutual correspondences between template and target and uses the Laplace-Beltrami operator $\mathbf{L}_\mathrm{r} \in \mathbb{R}^{p_\mathrm{X} \times p_\mathrm{X}}$ computed on the original shape, usually the reference shape (also called template shape) $\mathbf{X}_\mathrm{r} \in \mathbb{R}^{p_\mathrm{X} \times 3}$ as a regularization, controlled by the stiffness parameter $\lambda \in \mathbb{R}_{\geq 0}$. A higher λ puts more weight to the regularization term with the Laplace-Beltrami operator of the equation, leading to a mesh which retains its original shape (this is visualized in Fig. 4.2). For a low λ, the original template shape is disregarded and is mapped closer to the target mesh, which might lead to irregularities in the projection. This template projection step can be solved for the deformed shape $\mathbf{X} \in \mathbb{R}^{p_\mathrm{X} \times 3}$ using [33, 115]:

$$\begin{bmatrix} \lambda \mathbf{L}_\mathrm{r} \\ \mathbf{S}_\mathrm{X} \end{bmatrix} \mathbf{X} = \begin{bmatrix} \lambda \mathbf{L}_\mathrm{r} \mathbf{X}_\mathrm{r} \\ \mathbf{S}_\mathrm{Y} \mathbf{Y} \end{bmatrix} \tag{4.8}$$

The two Boolean selection matrices $\mathbf{S}_\mathrm{X} \in [0,1]^{k \times p_\mathrm{X}}$ and $\mathbf{S}_\mathrm{Y} \in [0,1]^{k \times p_\mathrm{Y}}$ select the k correspondences on the template (or reference shape) $\mathbf{X}_\mathrm{r}$ and target $\mathbf{Y}$. p_X denotes the number of template points, p_Y the number of target points.

4.4 Statistical Modeling

After correspondence establishment, the shapes in dense correspondence have to be aligned to remove translation and rotation using GPA (see Section 4.2.2). After this step, the actual statistical modeling can be performed.[3]

Each target in dense correspondence $\mathbf{\Gamma}_i^{\mathrm{c,a}}$ is regarded as a multivariate, independent observation of a random variable and its probability distribution. For this reason, each $\mathbf{\Gamma}_i^{\mathrm{c,a}}$ is vectorized into a 1-dimensional vector representing a multivariate observation of one shape $\mathbf{s}_i = [x_i^1, y_i^1, z_i^1, x_i^2, y_i^2, z_i^2, \dots x_i^{p_\mathrm{r}}, y_i^{p_\mathrm{r}}, z_i^{p_\mathrm{r}}]^\mathsf{T}$:

$$\mathrm{vec}(\mathbf{\Gamma}_i^{\mathrm{c,a}}) = \mathbf{s}_i \in \mathbb{R}^{3p_\mathrm{r}} \tag{4.9}$$

All N observations are arranged into a 2D observation or data matrix $\mathbf{S} \in \mathbb{R}^{3p_\mathrm{r} \times N}$ in which the full statistical information of the SSM is contained. The next step is to extract the relevant information from the matrix. SSMs typically model the shape variation as a multivariate Gaussian distribution, which can be entirely described by its mean shape $\boldsymbol{\mu} \in \mathbb{R}^{3p_\mathrm{r}}$ and its covariance $\mathbf{\Sigma} \in \mathbb{R}^{3p_\mathrm{r} \times 3p_\mathrm{r}}$. The mean shape can be subtracted from the observation matrix to obtain the zero-mean observation matrix:

$$\mathbf{S}_{\mathrm{zm}} = \mathbf{S} - \boldsymbol{\mu} \tag{4.10}$$

The zero-mean observation matrix is analyzed using dimension reduction such as PCA. Alternatives to ordinary PCA are probabilistic PCA [131] (suitable for including incomplete observations) and weighted principal component analysis (WPCA) [33] which allows to assign different weights to each point.

4.4.1 Weighted Principal Component Analysis

Applying ordinary PCA treats each point with the same weight, while WPCA enables the usage of weights for each points. For example, it might be desirable to use a denser point concentration for the face and ears in a head model since more detailed expression is expected or desired. The over-representation of the face can be balanced out using WPCA and the cranium is not under-represented [33].

For SSMs, there are usually many more points than observations ($p_\mathrm{r} >> N$), so WPCA can be computed efficiently using the weighted Gram matrix $\mathbf{G}_\mathrm{W} \in \mathbb{R}^{N \times N}$ (resulting in the decomposition of this matrix instead of one with the size of $\mathbb{R}^{3p_\mathrm{r} \times 3p_\mathrm{r}}$) [115]. The sparse weight matrix $\mathbf{W} \in \mathbb{R}^{3p_\mathrm{r} \times 3p_\mathrm{r}}$ assigns a weight to each

[3] If GPA is not performed, translations and rotations will be present in the first principal components of the model.

variable (point coordinate) and to each edge for connected vertices. The weighted Gram matrix $\mathbf{G}_{\mathrm{W}} \in \mathbb{R}^{N \times N}$ can be computed as

$$\mathbf{G}_{\mathrm{W}} = \mathbf{S}_{\mathrm{zm}}^{\mathsf{T}} \mathbf{W} \mathbf{S}_{\mathrm{zm}} \tag{4.11}$$

to perform the eigendecomposition of $\mathbf{G}_{\mathrm{W}}$ as

$$\mathbf{G}_{\mathrm{W}} = \mathbf{U}_{\mathrm{G}} \mathbf{\Lambda}_{\mathrm{G}} \mathbf{U}_{\mathrm{G}}^{\mathsf{T}}. \tag{4.12}$$

The principal components $\mathbf{V} \in \mathbb{R}^{3p_{\mathrm{r}} \times N}$ of the SSM can be computed as

$$\mathbf{V} = \mathbf{S}_{\mathrm{zm}} \mathbf{U}_{\mathrm{G}} \mathbf{\Lambda}_{\mathrm{G}}^{-\frac{1}{2}}, \tag{4.13}$$

The desired eigenvalues $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ of the sample covariance matrix can be re-scaled from the Gram matrix $\mathbf{G}_{\mathrm{W}}$:

$$\mathbf{\Lambda} = \frac{1}{N-1} \mathbf{\Lambda}_{\mathrm{G}}. \tag{4.14}$$

Each observation can be re-parameterized using the parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^{N}$ that assigns a weight to each principal component:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{V} \mathbf{\Lambda}^{\frac{1}{2}} \boldsymbol{\alpha} \tag{4.15}$$

This representation of an SSM is preferred over storing the full covariance matrix. Some of the advantages and applications include:

- Small storage requirements, even for large matrices due to the decomposition of the covariance matrix via PCA.
- The shape variance is concentrated in the first few components, so the final components contain mostly noise. Therefore, the first k principal components can be selected without losing much information. k is often determined as a ratio of the total variance in the dataset.[4]
- Drawing random shapes only requires drawing a random $\boldsymbol{\alpha}$ from a normal distribution.
- Fitting an SSM to unseen data requires only spatial alignment of an unseen sample $\mathbf{s}_{\mathrm{unseen}}$ with $\bar{\mathbf{s}}$ and solving of Eq. (4.15).
- Translation of probabilistic applications such as posterior probability into the shape space (e.g., posterior shape modeling [108]).

[4]The last principal component will always be $\mathbf{0}$, since the mean shape $\bar{\mathbf{s}}$ was subtracted before computing the covariance matrix.

4.5 Model Evaluation

Evaluation of statistical models can be subdivided into two categories: Evaluating the registration of each subject to the morphed reference using distance error metrics and evaluating the model itself using compactness, generalization, and specificity [132].

4.5.1 Registration Evaluation

Landmark points: One of the most important registration metrics [132] are landmark points, which are easily identifiable, corresponding points on both meshes. Landmark errors are computed as Euclidean distances. By virtue of its definition, those landmark points are sparse and morphologically consistent, but they do not necessarily represent the whole object well. For example, there are several facial landmarks, but they do not cover the back of the head.[5] While landmark errors indicate well if the global registration of the two meshes is good, they are "blind" to local registration.

Vertex-to-nearest-neighbors: Vertex-to-nearest-neighbor distances evaluate all points and are independent from any landmarks. For each point of the reference mesh (the term "point" and "vertex" are in this regard used interchangeably), the corresponding nearest neighbor on the target mesh is determined and their Euclidean distance is computed. However, the vertex-to-nearest-neighbor distance is low as long as the two surfaces are morphed close to each other, regardless of if the other surface is morphologically correct. For example, a nose can be morphed to a chin due to poor initial alignment, but the vertex-to-nearest-neighbor-error will be low. This metric is therefore suitable for local registration, but is "blind" to the global registration. A variant of this approach includes vertex-to-surface or point-to-plain (point distance to the surface). This is usually more accurate but requires face information and is computationally more expensive.

4.5.2 Shape Model Evaluation Metrics

Compactness Compactness describes the model's ability to contain much of the model's variance in the first few components. It is computed as the sum of the eigenvalues of the sample covariance matrix and is often normalized. Compactness is not an error metric and therefore higher values are considered better.

[5] If they were densely available and would cover the complete shape (the ideal case), there would not be a need to perform a registration in the first place.

Generalization Generalization describes the model's ability to generalize well to unseen data. It is computed using an leave-one-out approach in which the model is constructed using $N - 1$ samples and is fitted to the N^{th} sample. Generalization error is usually higher for the first components and becomes smaller the more components are used (because the left-out sample can be explained better if there are more components). A smaller generalization error is considered better.

Specificity Specificity is linked to data synthesis and describes the model's ability to create specific instances. Its computation requires to generate synthetic instances and the closest training sample is determined. Specificity is generally lower for the first principal components and increases with more components. A smaller specificity error is considered better.

PART II

DATA PREPARATION AND STATISTICAL MODELING

Chapter 5

Dataset and Preprocessing

5.1 3D Surface Scans and Landmarks

All data from this study were thankfully provided from the Department of Oral and Maxillofacial Surgery of the Heidelberg University Hospital, where all patients with craniofacial diseases are routinely recorded using a 3D surface imaging system (Canfield VECTRA-360-nine-pod, Canfield Science, Fairfield, NJ, USA) for monitoring and documentation purposes. Often, the patients were recorded multiple times before and after therapy, preoperatively to track the status of the disease and post-operatively to monitor and ensure correct skull development. The children wore tight-fitting hairnets to minimize artifacts caused by the hair. A standardized protocol was used, which had been examined and approved by the Ethics Committee Medical Faculty of the University of Heidelberg (ethics number S-237/2009). The study was carried out according to the Declaration of Helsinki and written informed consent was obtained from all parents. All scans were acquired between 2011 and 2021. For each recording, the scanner provided a triangular surface mesh which was later annotated with ten cephalometric landmarks and the medical diagnosis by clinical staff. The available landmarks for each scan are visualized in Fig. 5.1 and listed in Tab. 5.1.

5.2 Inclusion and Exclusion Criteria

For a classification study, the 3D patient geometry, landmarks for alignment, and the label with the clinical diagnosis were required. Additionally, most patients were recorded multiple times which could potentially introduce cross-over: If the same patient appeared in training and test set, the classifier could "cheat" by identifying patient-specific features, leading to an over-estimation of the classifier performance. Thus, for each patient only the preoperative scans closest to the operation date were selected and duplicate scans of the same patients were discarded.

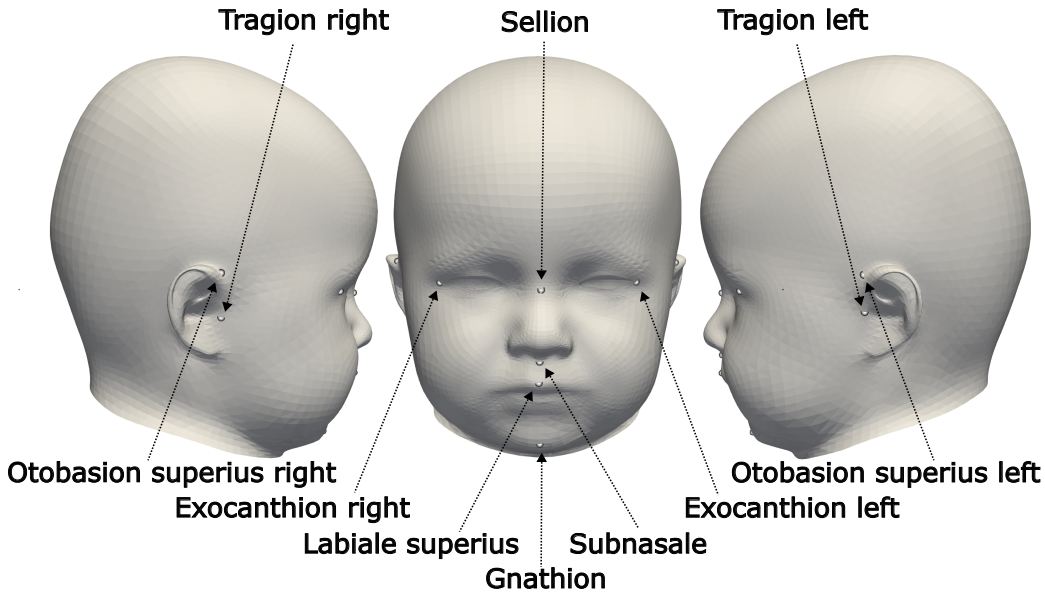


Figure 5.1: Landmarks provided in the dataset. Six out of ten landmarks were symmetric landmarks available on the left and right side. The same subject is shown from three perspectives.

Table 5.1: Landmarks on 3D surface scans provided by the medical staff. The cephalometric landmark notation of [133] was used.

Landmark	Abbreviation
Tragion (left and right)	(t_l) and (t_r)
Sellion	(se)
Exocanthion (left and right)	(ex_l) and (ex_r)
Subnasale	(sn)
Labiale Superius	(ls)
Otobasion superius (left and right)	(obs_l) and (obs_r)
Soft tissue gnathion	(gn)

While the initial dataset contained 7529 samples from 2553 different subjects, most of them had to be discarded beforehand: Many scans were duplicate scans from the same patients, did not have a patient age available or were older than 1.5 years. However, around 80 different labels were annotated as a ground truth pathology for those scans, but over 70 of them were duplicates or so rare they contained only one or two samples. Multi-suture synostostotic cases were removed. As 10-fold cross validation was planned, pathologies with fewer than ten instances were removed, including lamboid synostosis. An automatic pipeline was implemented using `pymeshlab` and `bash` to discard patients with corrupt scans, incomplete landmarks, or a missing 3D Slicer [134] transformation matrix to map the landmarks from the scanning device frame to real-world coordinates. Around 550 subjects

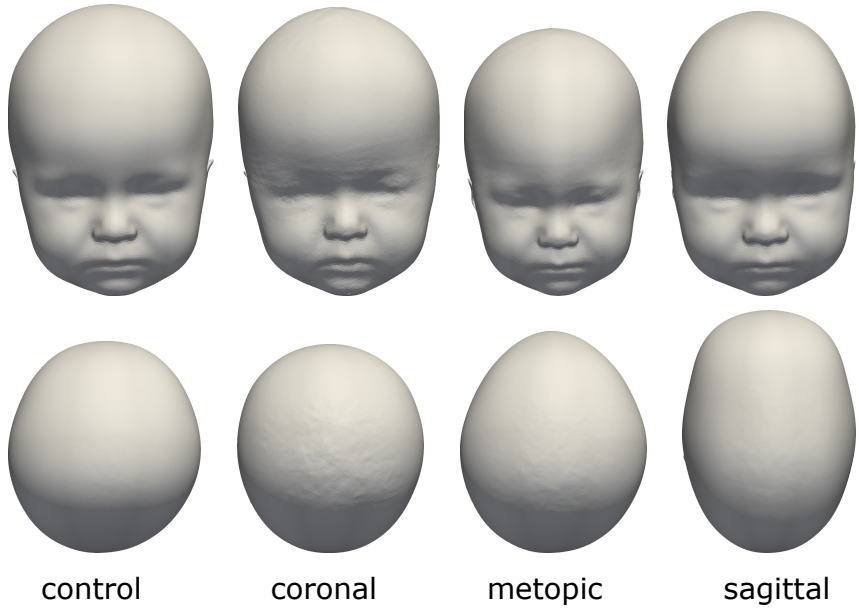


Figure 5.2: Head shapes of the four classes in the dataset. Top row: front view, bottom row: top view.

remained. As a last step, the remaining files were semi-automatically inspected from different perspectives and around 50 patients recordings had uncorrectable errors in their scans such as missing parts or large holes which also had to be removed. This resulted in the final dataset configuration of 496 subjects.

The following types of craniosynostosis patients had been selected: coronal (brachycephaly and unilateral anterior plagiocephaly), sagittal (scaphocephaly), or metopic suture fusion (trigonocephaly), as well as a control group without any suture fusion. The four classes are displayed in Fig. 5.2. Beside healthy subjects, the majority of the control group consisted of scans of children with positional plagiocephaly. While positional plagiocephaly patients were later treated with helmet therapy or laying repositioning, all craniosynostosis patients underwent surgical remodeling of the cranium. The final dataset consisted of 496 subjects. A violin plot [135–137] of the 496 patients' class and age distribution is displayed in Fig. 5.3. The distribution including left and right annotation as subclasses is displayed in Fig. 5.4. Regarding the selection of classes, this approach is comparable to other classification studies, which distinguished between craniosynostosis and non-craniosynostosis classes, in particular Mendoza et al. [29] and de Jong et al. [13].

5.3 Preprocessing and Artifact Removal

The Python module `pymeshlab` from the open-source software Meshlab [138] was used to preprocess the 3D surface scans. Isolated parts, duplicate faces and vertices were removed, and holes in the surface scans were closed in a fully automated

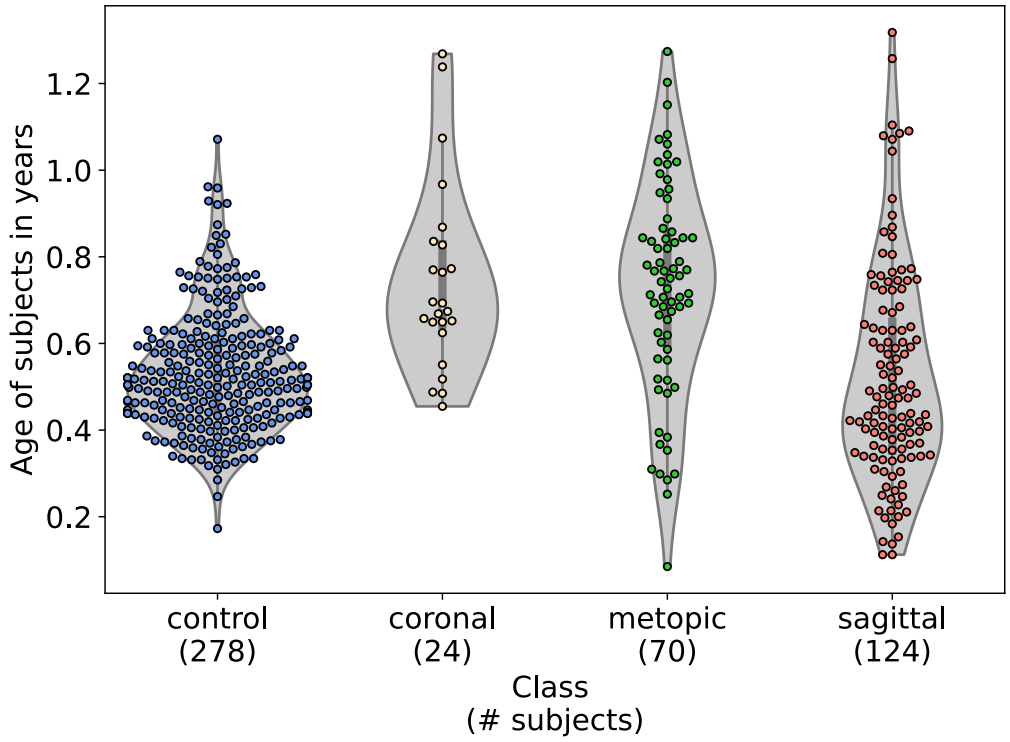


Figure 5.3: Violinplot of the class and age distribution of the subjects in the dataset. Parenthesis indicate number of samples per class.

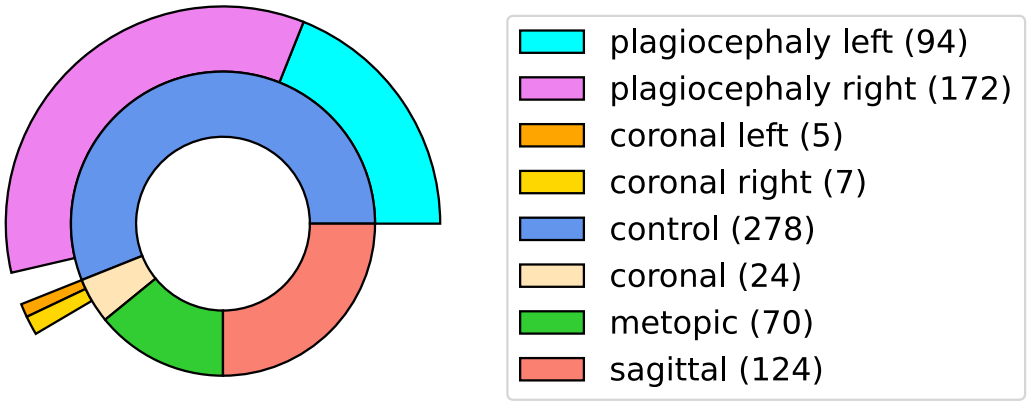


Figure 5.4: The class distribution of the dataset including subclasses with left and right annotations. Parenthesis indicate number of samples per class. If not specified, the subclass was ignored.

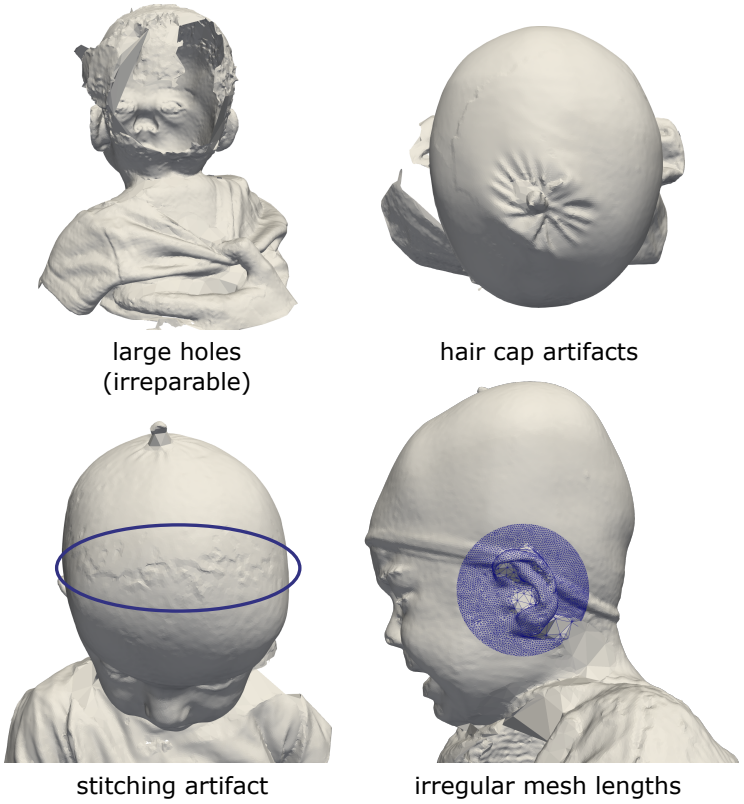


Figure 5.5: Repairable and non-repairable artifacts in the dataset. Large holes (top left) were non-repairable and made the scan unusable, hair cap artifacts (top right) were not removed and did not lead to poor performance. Stitching artifacts (bottom left) appeared like a scar and could be removed using re-meshing. Irregular mesh lengths (bottom right) were often close to the ears and could mostly be corrected using re-meshing, but were sometimes too large so that the scan had to be removed.

manner, as those types of artifacts could lead to incorrect data in the distance maps. Typical artifacts are depicted in Fig. 5.5. Additionally, parts of and around the ears were often characterized by large edge lengths, so isotropic explicit re-meshing [139] with a target length of 1 mm was used to obtain regular meshes. The medical staffs' clothes and hands could be ignored since they only had body contact at the torso of the child to position it and did not affect the scan of the head. Everything below the child's neck could be cut off to speed up computation during feature extraction, template morphing, or image creation.

5.4 Sellion-Tragion Orientation

The sellion tragion orientation (STO) is a coordinate frame defined during this thesis from the sellion and left and right tragion landmarks which enables a systematic extraction of shape parameters across different scans and will be used in Chapter 7 and Chapter 9. Its origin and axes were defined in Cartesian space using three

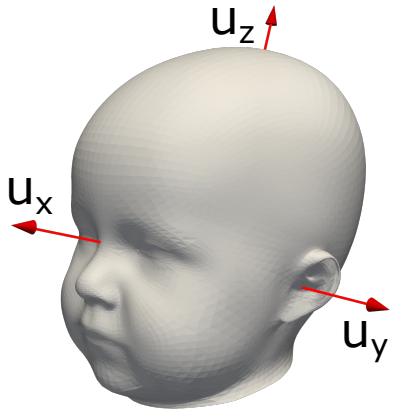


Figure 5.6: Sellion tragion orientation coordinate axes constructed from sellion, left tragion, and right tragion.

landmarks (left and right tragion, located on the ears, as well as the sellion, located on the nose) in a similar manner to the frontal, sagittal, and median axes commonly used in various medical disciplines and is depicted in Fig. 5.6.

The center point or origin $\mathbf{p}_c$ was defined as the midpoint of left and right tragion ($\mathbf{p}_{tl}$ and $\mathbf{p}_{tr}$):

$$\mathbf{p}_c = \frac{1}{2}\left(\mathbf{p}_{tl} + \mathbf{p}_{tr}\right) \tag{5.1}$$

The two landmarks were located on different ends, so the origin was approximately in the center of the head. The definition proposed here is reminiscent of the cranial focus point definition [35] for computed tomography (CT) data. The axis direction $\mathbf{u}_x$ (corresponding to the frontal axis) was defined as the direction from the origin to the sellion located on the nose:

$$\mathbf{u}_x = \mathbf{p}_s - \mathbf{p}_c \tag{5.2}$$

$\mathbf{u}_y$ (corresponding to the median axis) was defined orthogonal to $\mathbf{u}_x$ from the center to the left tragion:

$$\mathbf{u}_y = (\mathbf{p}_{tl} - \mathbf{p}_c) - \mathbf{u}_x \frac{\mathbf{u}_x \cdot (\mathbf{p}_{tl} - \mathbf{p}_c)}{||\mathbf{u}_x||}. \tag{5.3}$$

$\mathbf{u}_z$ was constructed to be orthogonal to the two previous directions, thus corresponding to the sagittal axis:

$$\mathbf{u}_z = \mathbf{u}_x \times \mathbf{u}_y \tag{5.4}$$

The direction vectors $[\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z]^T$ were each normalized to length 1 mm so that they created an orthonormal basis $[\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z]^T$.

Chapter 6

Statistical Craniosynostosis Head Model

This chapter quotes partly in verbatim from the related open-access publication licensed under CC-BY in Diagnostics [140]. The original publication comprises both the publicly available model [141] and the statistical shape model (SSM)-based classification approach based on the model, originally published with a reduced dataset of 367 samples. For this thesis, they were separated into two different chapters, one for the SSM (Chapter 6) and one for the classification approach (Chapter 8).

6.1 Introduction

As outlined in the introduction in Chapter 1, statistical modeling is a popular approach to synthesize synthetic, large-scale datasets with a high degree of individual representation. This is useful to reduce the dependency on clinical data, thus reducing costs and making data available to other research groups. For a general overview about statistical shape modeling, it is referred to Chapter 4. This introduction covers the more specific applications of SSMs to craniosynostosis.

For general-purpose head and face models [33, 110, 114], making the model publicly available to enable data synthesis is often a key aspect of the publication. In the field of craniosynostosis assessment, models are mostly used for statistical purposes or to compare pre-and post-operative craniosynostosis patients [33, 115, 142]. Data synthesis approaches have not been translated to SSMs of craniosynostosis patients despite the lack of publicly available datasets, which has been acknowledged [38, 39] and could boost also approaches using 2D photography data [143].

The goal of this work is therefore to create an SSM of the head containing the pathologic features of craniosynostosis patients, which can be made publicly available and used for data synthesis. As a later step (see Chapter 8), the model will be changed to a cranium model to enable an SSM-based classification approach.

6.2 Methods

6.2.1 Shape Model Creation

For a description of the dataset, the classes and the applied preprocessing to the data, the reader is referred to Chapter 5. The pipeline for the creation of the SSM employed in this chapter follows the structure described in the fundamentals in Chapter 4, consisting of alignment, correspondence establishment, and statistical modeling. Therefore, a template was aligned and morphed to each subject, from which the point identifiers could be used for the statistical modeling.

The mean shape of the Liverpool-York child head model [115] was used as the basis for the template. The Liverpool-York-model had been constructed as a symmetric model with physiological subjects, so the model was free of biases toward any particular pathology. However, since the model did not contain any eyes and mouth, those were added as additional vertices and triangular faces. The initial vertex order was left unchanged, which made it possible to incorporate the proposed model into the other (and vice versa using a posterior modeling approach [108]). The final template had a mean edge length of 2.91 mm and $p_{\mathrm{H}} = 13151$ vertices. In order to increase expression of the SSM, all original $N = 496$ subjects were mirrored on the sagittal plane to increase the dataset size for the SSM to $2N = 2 \cdot 496 = 992$ subjects. Alignment and shape morphing was performed individually.

Spatial alignment For alignment, Procrustes analysis was used (see Section 4.2.1 in the fundamentals) which was applied on the landmarks of template and target. As the Liverpool-York child head model was constructed using children from two to 15 years, ordinary Procrustes analysis including scaling was chosen for alignment. This yielded a transformation to translate, rotate, and re-scale computed on all ten template and target landmarks. This transformation was applied to the whole template mesh.

Correspondence establishment For correspondence establishment among the template and all the subjects (see Section 4.3), four morphing methods which had already been applied successfully to head morphing by other groups were selected and employed: Laplace-Beltrami regularized projection (LBRP) [115], iterative coherent point drift (ICPD) [33], nonrigid iterative closest points affine (ICPA) [120], and nonrigid iterative closest point translation (ICPT) [120] were analyzed and employed during this thesis. The mathematical formulations of the LBRP morphing can be found in Section 4.3.2, while all hyper-parameters and the description of the ICPD, the ICPA and ICPT methods can be found in the appendix in Chapter A. For the LBRP, the morphing approach was divided in two steps, the first step with a high

regularization $\lambda = 10$ and a second step with a low regularization $\lambda = 0.1$ to let the template deform closer toward the target scan.

Statistical modeling Statistical modeling consisted of the pipeline described in Section 4.4: Rigid generalized Procrustes analysis (GPA) was applied to remove the non-shape related attributes translation and rotation from the morphed templates (see Section 4.2.2). Scale was considered an attribute of shape because features related to craniosynostosis could depend on the patient's age and head size.

For statistical modeling, weighted principal component analysis (WPCA) as described in Section 4.4.1 was used instead of ordinary principal component analysis (PCA) since ordinary PCA would have resulted in the majority of variance in the facial parts of the head model since there was the majority of vertices. The weights for each point were assigned according to the mass matrix $\mathbf{M} \in \mathbb{R}^{p_\mathrm{H} \times p_\mathrm{H}}$, which was composed of per-vertex weights and per-edge weights in a similar manner to barycentric cells: The diagonal elements of $\mathbf{M}$ represented the vertex weights. Each vertex weight was defined as the sum of the area of the adjacent faces for which this vertex was the nearest neighbor. Likewise, the non-diagonal elements represented the edge weights and each edge weight was defined as the sum of the area of the adjacent faces for which this edge was the closest edge. To account for the vectorized representation of the observations, the mass matrix was stretched by factor 3 and nearest-neighbor-interpolated, resulting in $\mathbf{M}_3 \in \mathbb{R}^{3p_\mathrm{H} \times 3p_\mathrm{H}}$. The computation of the Gram matrix and performing the actual singular value decomposition (SVD) is described in Section 4.4.1. This resulted in the typical representation of the SSM:

$$\mathbf{s}_\mathrm{H} = \bar{\mathbf{s}}_\mathbf{H} + \mathbf{V}_\mathrm{H} \mathbf{\Lambda}_\mathrm{H}^{\frac{1}{2}} \boldsymbol{\alpha}_\mathrm{H} \tag{6.1}$$

$\bar{\mathbf{s}}_\mathrm{H} \in \mathbb{R}^{3p_\mathrm{H}}$ denoted the vectorized mean shape, $\mathbf{V}_\mathrm{H} \in \mathbb{R}^{3p_\mathrm{H} \times 2N}$ denoted the principal components of the SSM, $\mathbf{\Lambda}_\mathrm{H} \in \mathbb{R}^{2N \times 2N}$ the eigenvalues of the sample covariance matrix, and $\boldsymbol{\alpha}_\mathrm{H} \in \mathbb{R}^{2N}$ the shape parameter vector which was computed on the augmented dataset consisting of the original and the mirrored samples of the full head.

Overall, one SSM of the full head was created, as well as four submodels of each class control, coronal, metopic, and sagittal. A modified version of the full head model would be relevant for the classification approach in Chapter 8, while the four submodels of each class were suitable for the generation of synthetic samples, required in Chapter 10. A texture model was also created to provide texture for 2D image-rendering applications.

6.2.2 Texture Model

The texture model was created across all subjects since no distinctions between the subclasses were expected. The SSM used a vectorized representation of all points as:

$$\mathbf{s} = [x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_{p_\mathrm{H}}, y_{p_\mathrm{H}}, z_{p_\mathrm{H}}]^\mathsf{T} \tag{6.2}$$

The texture model used therefore a vectorized representation of all color values:

$$\mathbf{t} = [r_1, g_1, b_1, r_2, g_2, b_2, \ldots, r_{p_\mathrm{H}}, g_{p_\mathrm{H}}, b_{p_\mathrm{H}}]^\mathsf{T} \tag{6.3}$$

For each point in correspondence, the color value was nearest-neighbor-mapped from the texture of the target mesh to the morphed template.

6.2.3 Pathology Change

To demonstrate the possibilities of the SSM, an exemplary application is presented: The head of a scaphocephaly patient is changed toward the control group, as shown in Fig. 6.7. This is a variation of changing a labeled attribution [110] (other examples include gender or weight) using linear regression. The pathology of an individual sample can be changed as

$$\boldsymbol{\alpha}_{\mathrm{ID,control}} = \boldsymbol{\alpha}_{\mathrm{ID}} + \overline{\boldsymbol{\alpha}}_{\mathrm{control}} - \overline{\boldsymbol{\alpha}}_{\mathrm{sagittal}} \tag{6.4}$$

with $\bar{\boldsymbol{\alpha}}_{\mathrm{class}}$ denoting the mean parameter vector of a specific class, $\boldsymbol{\alpha}_{\mathrm{ID}}$ the parameter vector from a specific patient, and $\boldsymbol{\alpha}_{\mathrm{ID,class}}$ the parameter vector of the subject with a specific class attribute added.

6.3 Results

6.3.1 Shape Model Publication

During the creation of this thesis, an SSM was made publicly available. The SSM was created from an earlier version of this dataset which contained 367 subjects and was published on Zenodo[1] [141]. The model contained the SSM, a texture model, triangular cell information, the class-specific submodels, and 100 instances of each model sampled from a Gaussian distribution. For more information, it is referred to the relevant publication [140].[2]

[1] https://zenodo.org/record/6390158

[2] Since the publicly available dataset [141] is linked to the publication [140], the dataset was not updated to the current model with 496 subjects. The principal components did not change substantially, while generalization error decreased by 2 mm and specificity error depending on the morphing approach by ≈ 1 mm.

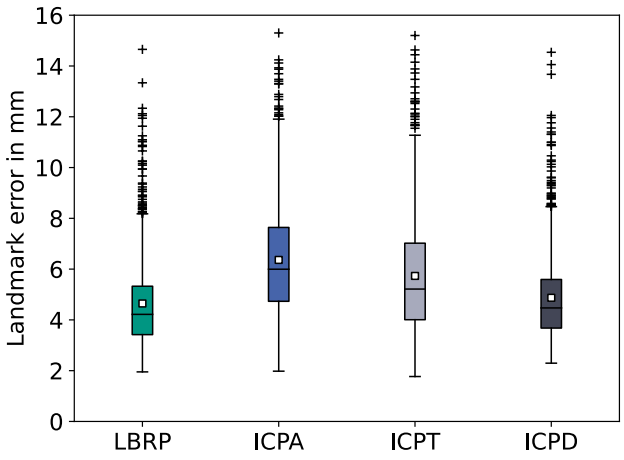


Figure 6.1: Boxplot of landmark errors for the four morphing approaches.

6.3.2 Morphing and Shape Model Evaluation

The template morphing approaches were evaluated using two metrics: landmark errors and vertex-to-nearest-neighbor distances. Landmark errors provide sparse point-to-point errors on known correspondences. Vertex-to-nearest-neighbor distances evaluated how close the template had been fitted onto the target points without taking into account if the nearest neighbor was morphologically correct. Fig. 6.1 shows landmark errors for the four morphing algorithms and reveals that the LBRP and ICPD methods scored lower landmark errors than ICPA and ICPT. Fig. 6.2 shows the mean vertex-to-nearest-neighbor-distances and reveals reverse trends compared to Fig. 6.1. ICPA showed the closest fit to the surface while LBRP showed that 10 % of the subjects had errors larger than 1 mm.

For shape model evaluation the three metrics compactness, generalization, and specificity [132, 144] were used. Compactness determines the model's ability to capture most of the variance with few components, generalization the model's ability to fit to unknown observations, and specificity the model's ability to create synthetic instances similar to the training data. Compactness, generalization, and specificity are presented in Fig. 6.3. The most compact models across all components were ICPA and ICPT which also had larger specificity errors. Generalization errors were lowest for LBRP and ICPT, but overall below 2 mm when using more than 10 model components. LBRP was chosen as the final morphing method as it scored best two of the five metrics (landmark error and specificity error), and scored the best generalization up to 40 components.

For qualitative comparison, the first model components of the LBRP method are depicted in Fig. 6.4. The first principal component changed primarily size. The second component affected both pathology and size, mostly shrinking the head

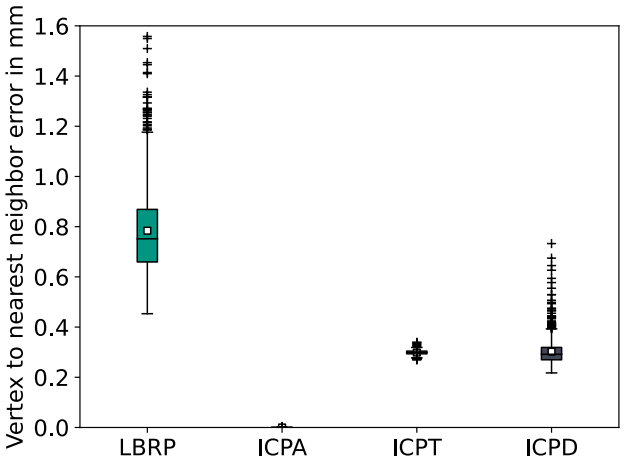


Figure 6.2: Boxplot of vertex-to-nearest-neighbor distances for the four morphing approaches.

to a shape reminiscent of a sagittal synostosis in negative direction and increasing the cranial size in positive direction similar to a coronal suture fusion. The third component influenced pathologies related to metopic and sagittal suture fusion. The mean shape submodels are presented in Fig. 6.5 and depict the expected pathologies clearly. The first three principal components of the texture model are depicted in Fig. 6.6. The first component changed brightness from dark to bright, the second component influenced mostly the color of the hair cap, and the last component had an effect on skin color and slightly influenced the hair cap color.

The pathology translation and attribution change from a scaphocephaly patient to the control group is visualized in Fig. 6.7. The individual could still be recognized as the same person, but the head shape had changed.

On a 3.7 GHz 6-Core Intel i5 processor, the computation times for the whole dataset of $2N = 496 \cdot 2$ patients for LBRP, ICPT, ICPA, and ICPD were 2 h 12 min, 50 h 24 min, 59 h 20 min, and 429 h 20 min. The LBRP was therefore around 20-fold less time-consuming than ICPT and ICPA, and around 200-fold less time-consuming than the ICPD approach. Real-time capability might not be an issue for a one-shot creation of an SSM as a publicly available dataset, but it would not be scalable if used in clinical practice (i.e., including an additional recording to an already existing model).

6.4 Discussion

The SSM united statistical information of 496 subjects and their mirrored twins with and without craniosynostosis. To date, many methods presented by various authors rely on in-house datasets making quantitative comparisons difficult. A set

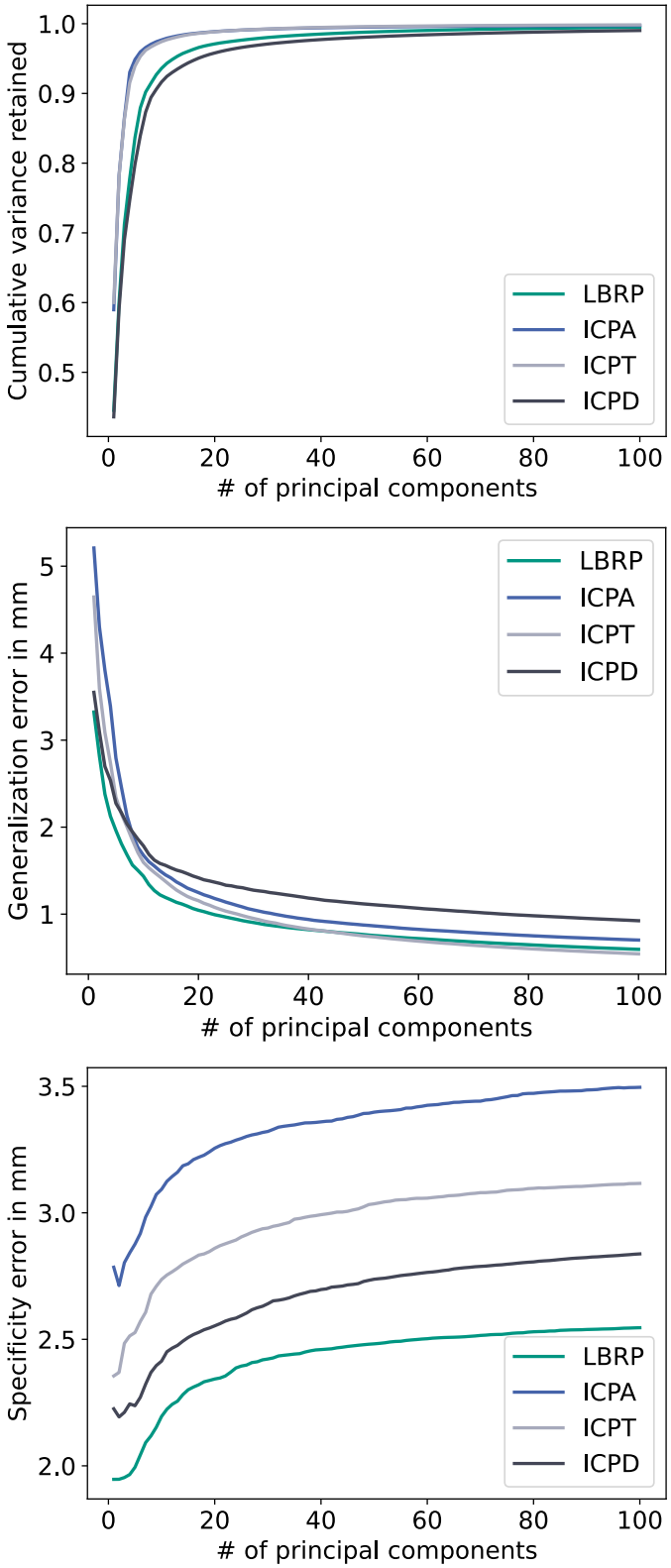


Figure 6.3: Compactness, generalization, and specificty of the four shape models depending on the morphing approach. For compactness, higher means better, for generalization and specificity, lower means better.

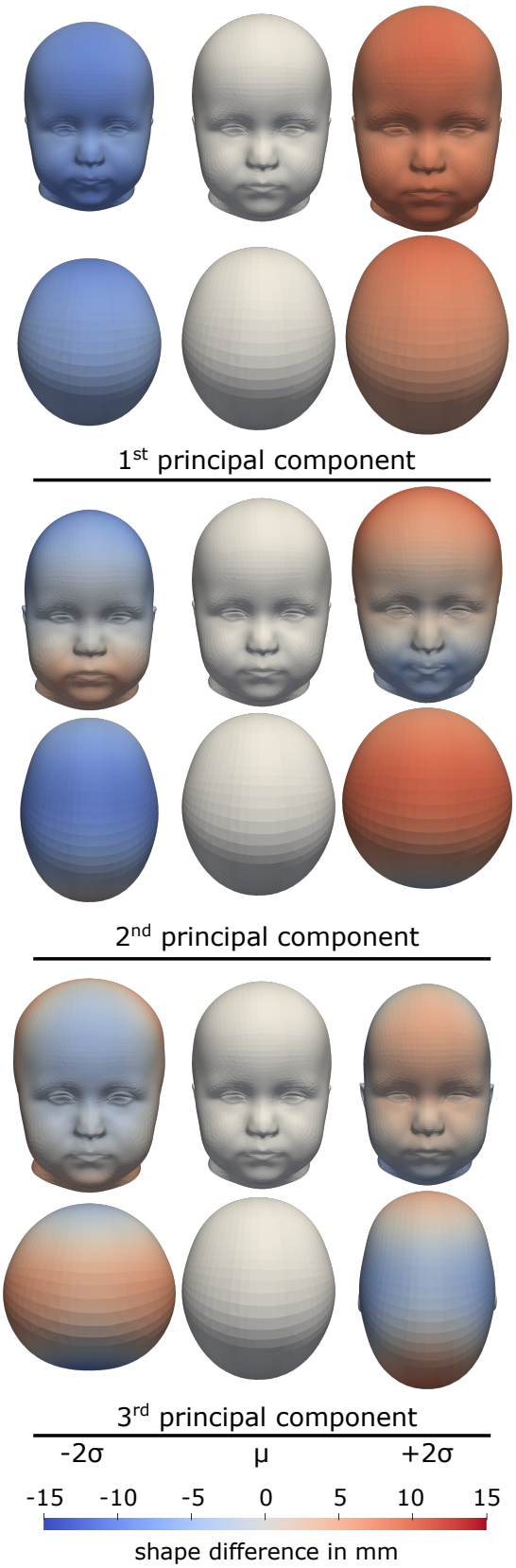


Figure 6.4: Principal components of the full statistical shape model constructed using the LBRP method, the colorbar indicates the Euclidean differences to the mean shape.

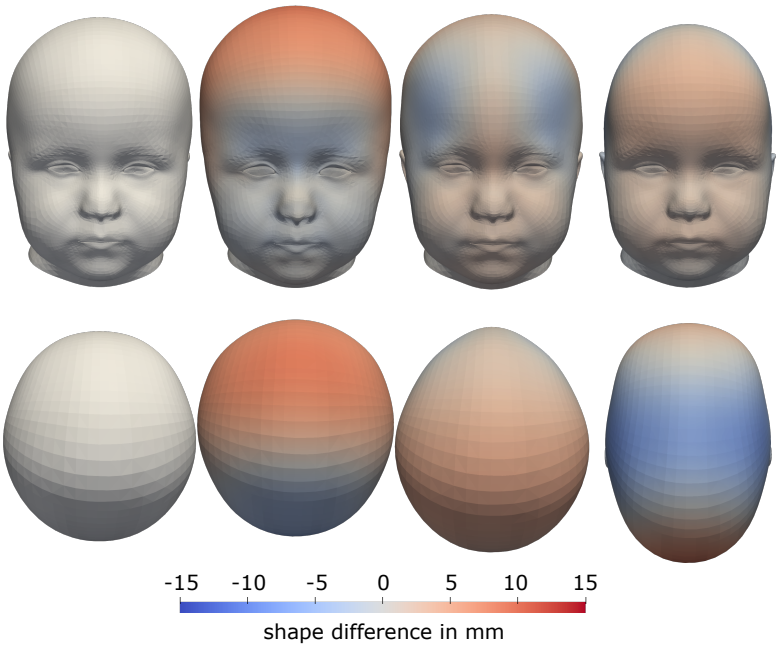


Figure 6.5: Mean shapes of the submodels constructed with the LBRP method, the colorbar indicates the Euclidean differences to the mean shape of the control model.

of synthetic 3D surface head scans of the SSM could help creating a large patient cohort for a reproducible evaluation of methods to assess craniosynostosis.

The principal components reflected the pathologies present in the dataset and showed realistic results. A comparison with other craniosynostosis-related SSMs is difficult since there are no publicly available models by other groups. In the medical field, studies which created shape models for craniosynostosis [31, 145] did not include quantitative metrics such as landmark error, compactness, generalization, and specificity. The most comparable SSM might be the Liverpool-York-Model [33, 115], as it is a full head model and also contained a submodel comprising children from 2 to 15 years. Compared to the Liverpool-York head model [33], the LBRP model employed in this thesis has similar landmark and vertex-to-vertex errors, but a higher compactness and lower generalization and specificity errors. However, the comparison has to be taken with a grain of salt since a substantially smaller dataset with different samples and age profiles was used in this work. Neither of the models is qualitatively better or worse, but it can be argued that the proposed SSM model of craniosynostosis patients performs similar to state-of-the-art head models.

Multiple morphing methods were tested during this work, but no method was clearly superior. While the ICPA method excelled in terms of vertex-to-nearest-neighbor errors, the LBRP model had the lowest landmark errors, specificity errors, low generalization errors, and was the fastest morphing method. For this reason, it was selected as the morphing method to be further used during this thesis.

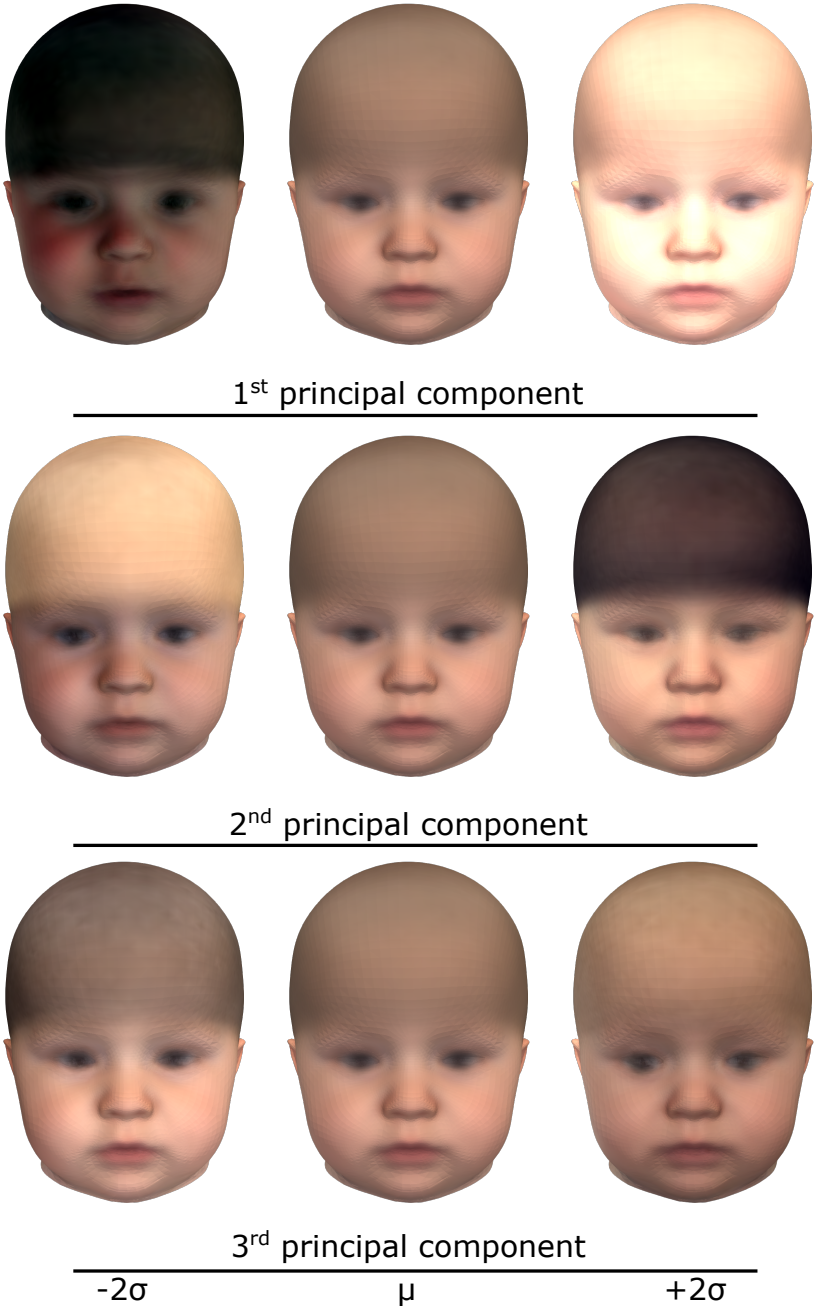


Figure 6.6: The texture model which can be employed for patient counseling. From top to bottom the first three principal components at $\sigma = 2$.

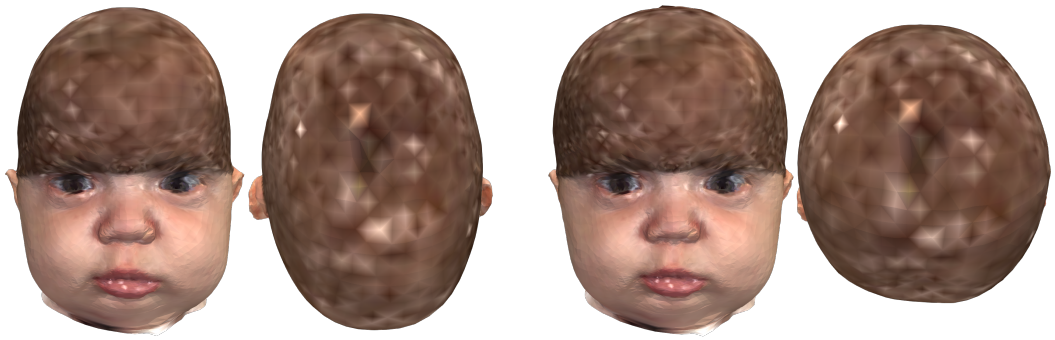


Figure 6.7: Patient pathology assessment using pathology change. Left: the original head shape of the scaphocephaly patient (front and top view). Right: the patient's head with removed pathology using the full SSM (front and top view). The original texture had been used which explains the noise.

Limitations of the model includes the control class of this study which was assembled by the scans of children who visited the hospital without indication to be treated surgically. This includes patients who were diagnosed being healthy and patients who were diagnosed having positional plagiocephaly. Thus, the control model represents a mixed group of children and should be used with caution when generating healthy subjects.

After a thorough research of the available literature, the model created in this study is the largest SSM of craniosynostosis patients and infants in general and the initial model is still the only one which has been made publicly available as of September 2023. Compared with the initially published model [140], the third principal component resulted in different shape changes, but otherwise, the metrics were qualitatively similar.

6.5 Conclusion

An SSM model creation pipeline for craniosynostosis patients was developed, suitable to create a state-of-the-art SSM of the head according to qualitative and quantitative metrics. The first publicly available SSM of craniosynostosis patients was derived using part of this dataset and the proposed methods. An approach for visualization and patient counseling was proposed and qualitatively evaluated. Researchers without access to clinical data can use the model for the assessment of head deformities.

PART III

CLASSIFICATION METHODS FOR CRANIOSYNOSTOSIS

Chapter 7

Classification Using Cephalometric Measurements

7.1 Introduction

Machine learning (ML) interpretability is a heavily debated issue in the scientific community of biomedical engineering. There are two approaches to ML model interpretability or often called "explainable artificial intelligence": The first option is trying to *explain* black box models, the second option is to use inherently explainable white box models in the first place. Many scientists in engineering [42] and institutions propose the usage of white box models over interpreting black box models to avoid discrimination and give users the "right to explanation" [43]. Typically cited examples for white box classification are decision trees (DTs) and k-nearest-neighbors (kNN) classifiers. The classification decision of DTs consists of simple Boolean if-else statements and can be fully retraced by visualizing the tree. However, if the DT's depth is too high or a random forest (RF) with many trees is used, the resulting visualization might be so complex, that it de facto ceases to be an easily understandable classifier for humans. kNN-classification can be visualized by retrieving the closest sample in the training set and the sample is assigned to the class in which there are the most neighbors in a given high-dimensional radius.

Ideally, an explainable white box classifier relies on already familiar clinical parameters and can be applied fast, easily, and economically. For this reason, the currently well established quantitative parameters cephalic index (CI) and cranial vault asymmetry index (CVAI) for assessing head shape are an apparent starting point for further analysis. However, the results of CI and CVAI values can vary depending on the extracted measurement position [146], and can limit measurement comparability [23]. Alternatively, it has been suggested to measure CI values of multiple heights during the assessment of sagittal synostosis on computed

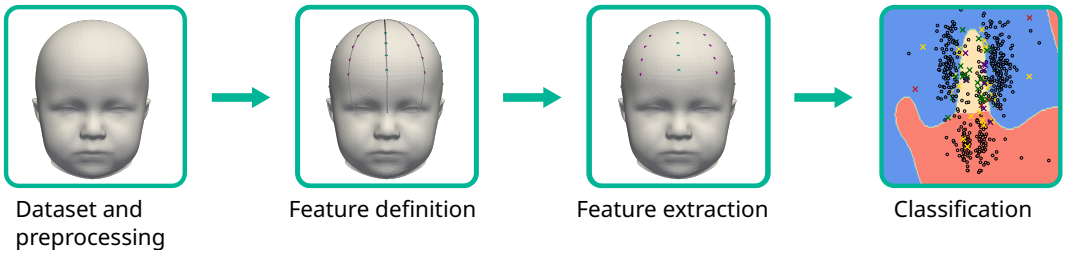


Figure 7.1: Schematic of the multi-height classification approach. After preprocessing of the dataset, the CI and CVAI could be extracted. The extraction of all values before the feature definition provides more flexibility in the feature definition to obtain the most relevant features before the classification.

tomography (CT) images [23, 147]. Focused on classifying head deformities on 3D surface scans instead of using CT scans, this work aims to make two contributions:

- **Single-height analysis:** The influence of the measurement height for the computed values of CI and CVAI, and their influence on classification performance will be systematically assessed and analyzed.
- **Multi-height classification:** Multi-height feature extraction approaches are proposed and tested with multiple classifiers.

White box classifiers such as DT, RF, and kNN are explicitly included and compared against other classification approaches in both the single-height domain and the multi-height domain.

7.2 Methods

The dataset and preprocessing steps used in this chapter are described in Chapter 5, resulting in 496 samples of four classes: control, coronal suture fusion, metopic suture fusion, and sagittal suture fusion. Fig. 7.1 visualizes the study schematic: CI and CVAI values were first computed, and feature extraction was performed in multiple ways. All classification experiments were subject to the same, stratified 10-fold cross-validation scheme with reproducible splits and a fixed random noise generator.

7.2.1 Clinical Parameter Extraction

The first step of the pipeline consisted of extracting the featues width w, length l, and diagonals $d_{-30°}$ and $d_{30°}$ for the later computation of the CI and the CVAI. This required a common frame of reference since the values were derived from rule-based algorithms that required angles and height values (see Section 2.3). The sellion tragion orientation (STO) (see Section 5.4) was used as a common frame

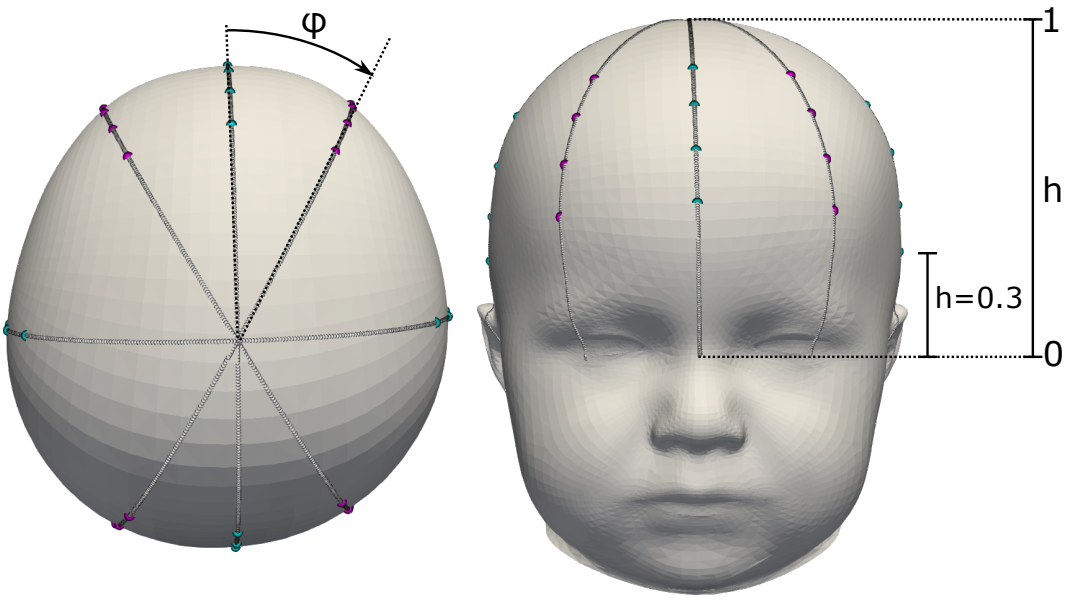


Figure 7.2: Intersection points of CI and CVAI values. The full height interval of $h = [0, 1)$ in steps of $\Delta h = 0.01$ is displayed. The Center-Steps extraction points are visualized for CI in cyan and CVAI points in magenta (mind the perspective distortion).

of reference across all subjects to compute CI and CVAIs values. A cylindrical coordinate system was employed for ray-casting with the radius r, and angle φ, and normalized height $h \in [0, 1)$, presented in Fig. 7.2. This allowed to compute CI and CVAI values dependent on height, by using width $w(h)$, length $l(h)$, and diagonals $d_{-30°}(h)$ and $d_{30°}(h)$, which had to be determined as the distance between intersection points on opposite ends. The intersection points were determined using ray-casting which was implemented using triangular ray intersection and the `vtk` python module [148] and is visualized in Fig. 7.2. Missing values for scans with holes or artifacts were interpolated using 1d linear interpolation in height direction.

7.2.2 Single-Height Analysis

As a first assessment, CI and CVAI were computed for later visualization of the height dependency:

$$\mathrm{CI}(h) = \frac{w(h)}{l(h)} \tag{7.1}$$

$$\mathrm{CVAI}(h) = \frac{d_{-30°}(h) - d_{30°}(h)}{\max(d_{-30°}(h), d_{30°}(h))} \tag{7.2}$$

As a second assessment, classification performance was determined as a function of the height. For this purpose, the classifiers had to be trained and evaluated for

Table 7.1: Multi-height classification approaches employed in this work. Single-height methods are in white, multi-height in light yellow.

Feature selection	Comment
Full-Heights	naïvely using very close values in $h \in [0, 1)$ with $\Delta h = 0.01$
Steps	from $h \in [0.1, 0.9]$ with a step-width of $\Delta h = 0.1$
Center-Steps	from $h \in [0.3, 0.6]$ with a step-width of $\Delta h = 0.1$
SHAP	mean absolute SHAP values
Middle	$h = 0.5$, single-height
Maximum	single-height features but not on a fixed point

each of the required heights, which was again sampled in the interval $h \in [0, 1)$ with $\Delta h = 0.01$.

The classification approach was carried out using a stratified 10-fold cross validation scheme and multiple classifiers were tested: linear discriminant analysis (LDA), support vector machine (SVM) with a linear kernel, naïve Bayes (NB), DT, RF, and kNN with $k = 5$. For the SVM, a one-versus-one with six binary linear SVMs was chosen to enable multi-class classification, for DT and the RF, the maximal depth was capped to six which acted as a regularization and prevented over-fitting. A single feature vector consisted of width w, length l, and diagonals d_{-30° and d_{30°. Before feeding the feature vector to the classifiers, all extracted features were divided by their Frobenius norm for normalization. For each height, all classifiers were trained and evaluated, yielding a total of $100 \cdot 6 \cdot 10 = 6000$ classification runs (100 height steps, 6 classifiers, and 10 cross-validation runs).

7.2.3 Multi-layer Classification

Multiple methods were designed to explore the classification of craniosynostosis on cephalometric parameters with multiple height measurements. They are summarized in Tab. 7.1 and were designed to have differences in the number and distribution of features.

The baseline consisted of the two single-height methods “Middle” and “Max” which used the values $h = 0.5$ and the maximum value. The four multi-height methods should be compared against the baseline: The “Full-Heights” method was designed as a naïve method using the full range of all values. The “Center-Steps” and “Steps” methods (see Tab. 7.1) were intended to find a trade-off between head coverage and reducing measurements. This trade-off between head coverage, measurement robustness and measurement effort avoided disturbances such as the eye cavities and difficult plane determination close to the tip of the head. The points of the “Center-Step” features are depicted in Fig. 7.2. The final feature approach used SHapley Additive exPlanations (SHAP) values [149]. SHAP values try to extract the most important features by varying the individual features in their inputs.

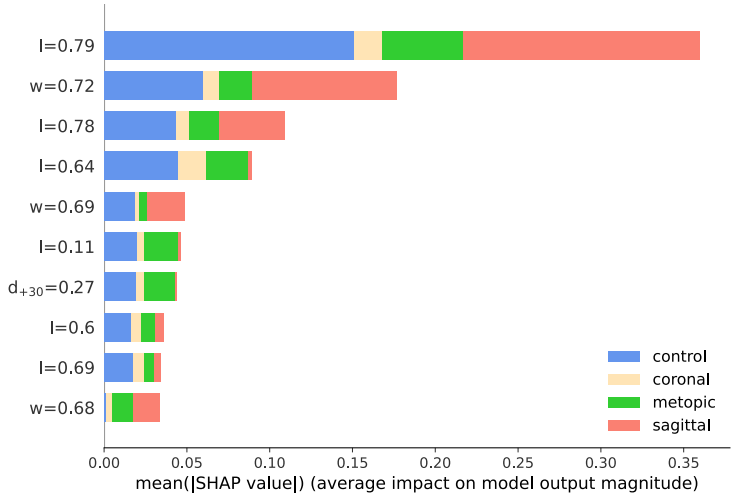


Figure 7.3: SHapley Additive exPlanations (SHAP) analysis of CI and CVAI in the dataset. The SHAP values were obtained by accumulating the feature points according to their DT feature contribution on the test sets of ten-fold cross-validation in an initial experiment before the final study.

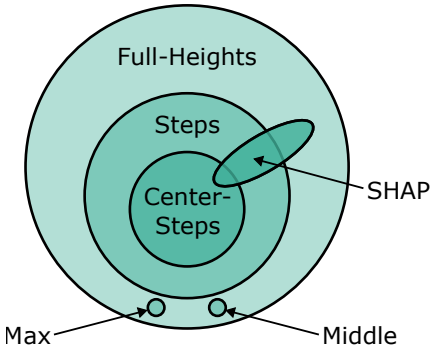


Figure 7.4: Overlap of CI and CVAI points displayed as sets of a Venn diagram. The SHAP features used values from the full head while the other multi-height features reduced the number of points from the full head step by step.

The SHAP values were selected from all extracted heights and were determined according to the maximum value of the mean absolute average importance by the tree explainer from the Python `shap` module. The SHAP values are presented in Fig. 7.3. The Venn diagram in Fig. 7.4 shows the relationships between the six different feature sets.

Multi-height classification was performed as described in the previous section (Section 7.2.2) with the same classifiers and the same training scheme: 10-fold stratified cross validation with consistent splits across all classifiers and the same seed for the random noise generator on $6 \cdot 6 \cdot 10 = 360$ classification runs (6 feature scenarios, 6 classifiers, 10 cross-validation runs). All scenario were evaluated using accuracy, G-mean, and macro F1-score.

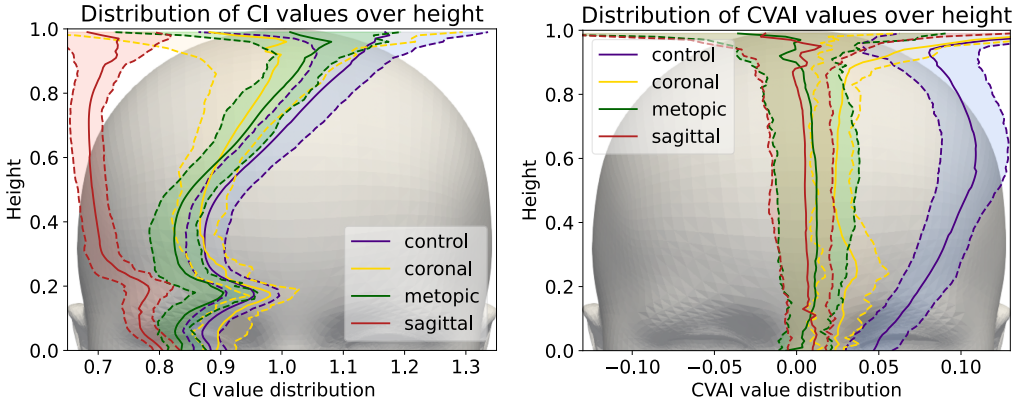


Figure 7.5: Distribution of cephalic index (CI) and cranial vault asymmetry index (CVAI) values across the subjects with respect to the extraction height. The background visualizes the values on the extracted height on the mean shape of the control group, i.e., the extracted height is aligned with the background and the resulting distribution of CI and CVAI is placed on the x-axis. The mean value is shown as a solid line, the 25^{th} and 75^{th} percentiles are shown as dotted lines for each class.

7.3 Results

7.3.1 Height Dependency Analysis

The resulting CI and CVAI values are displayed in Fig. 7.5. CVAI values were plotted as absolute values to prevent the plagiocephaly-biased split of the control group into two different groups. Alternative statistics (including classification results) explicitly defining the plagiocephaly groups are presented in the appendix in Chapter B.

Several observations could be made:

- **Divergence for $h > 0.9$:** Both CI and CVAI diverge with increasing height and above $h > 0.9$ as the measurement values become less robust.
- **CI notch:** The CI values showed a "notch" around $h = 0.2$. According to Fig. 7.5, this was close to the orbit (eye cavity) and the ear. This was an artifact and the values should therefore only be considered correct above $h = 0.3$.
- **CI values:** Except for the sagittal synostosis, CI values were larger the higher they were extracted. The most robust values could be observed at the height from $h \in [0.3, 0.5]$.
- **CVAI values:** CVAI values were mostly constant except for the control group which contained many cases of plagiocephaly. Plagiocephaly values peaked in $h \in [0.6, 0.7]$.

Regarding the single-height classification, Fig. 7.6 shows that the accuracy increased with the relative height and peaked for most classifiers at approximately $h = 0.4$. Fig. 7.6 shows accuracy, F1-score, and G-mean for classifiers trained on

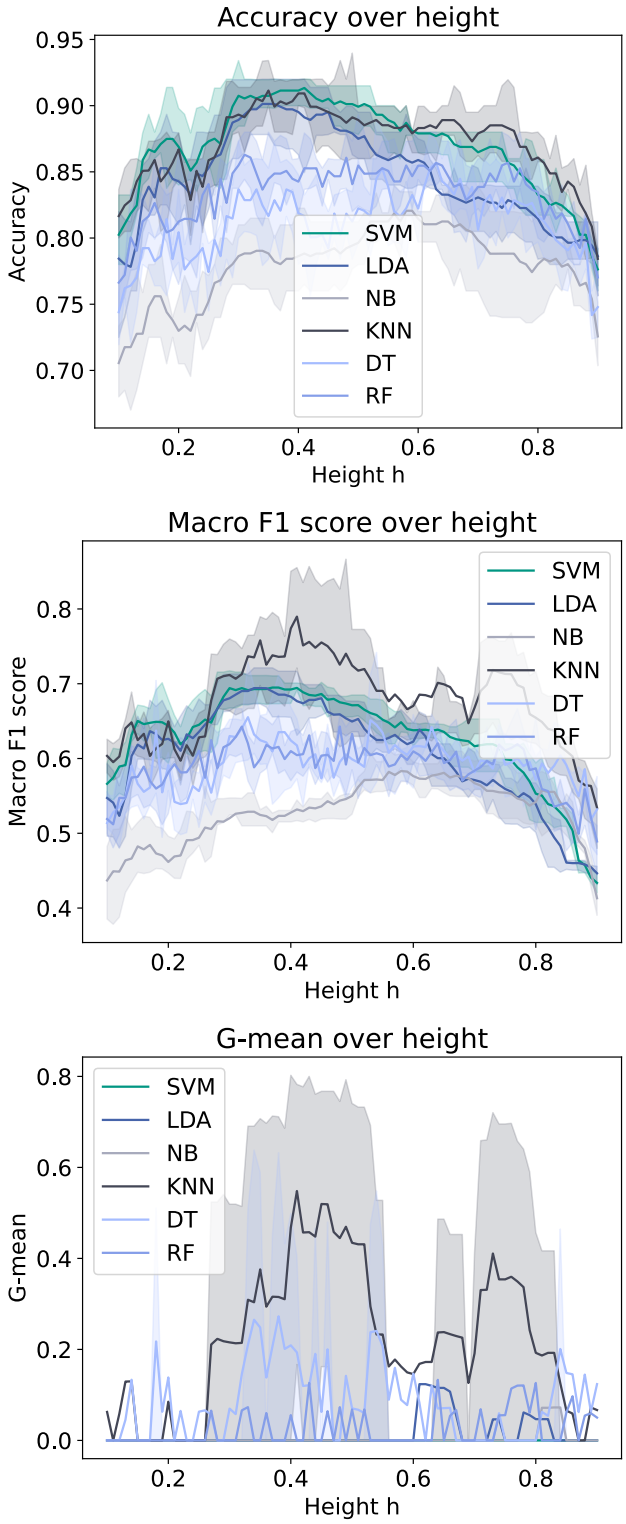


Figure 7.6: Classification performance of clinical parameters classifiers over the extraction height of CI and CVAI. The orbit artifact at $h = 0.2$ lead to a decrease in classification performance. Otherwise, classification performance increased until it reached a maximum between $h \in [0.6, 0.8]$.

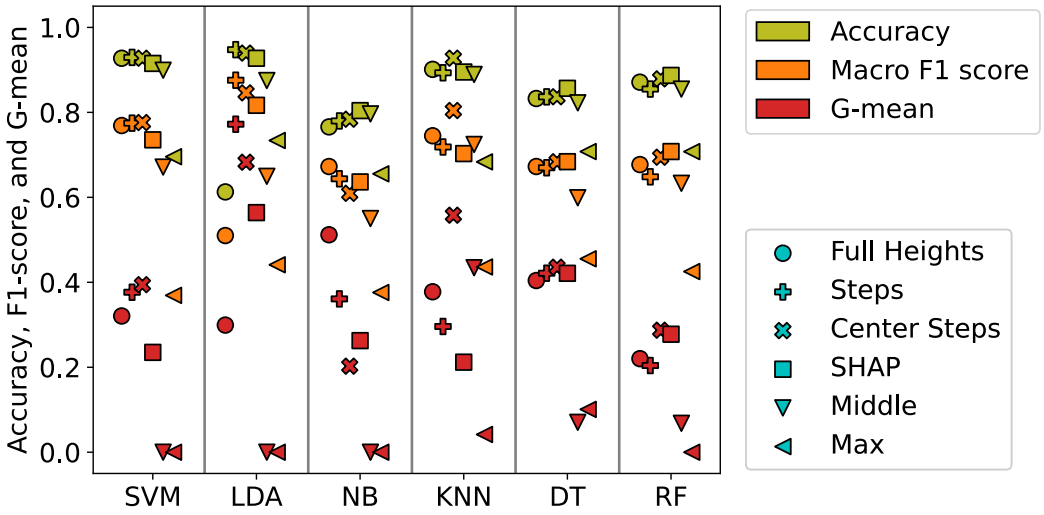


Figure 7.7: Mean classification performance for different classification approaches and different classifiers using 10-fold cross-validation. Each column shows a different classifier, colors show different mean performance metrics, while the symbols show different features.

different heights. For G-mean, only the kNN obtained results of 0.4 or higher, while SVM, LDA, and NB classifiers almost permanently remained at 0. The reason was that one of the classes (which turned out to be the least present class, coronal suture fusion) was never classified correctly in a single cross-validation run. Best F1-scores were mostly achieved in the interval of $h \in [0.4, 0.5]$ by the kNN classifier. Multiple classifiers such as SVM and NB score consistently a G-mean of 0 despite large accuracy values (for the SVM even larger than 0.9), indicating that accuracy alone is misleading to consider classification performance.

7.3.2 Multi-Height Classification

Fig. 7.7 shows the results for all metrics, classifiers, and features (the same results can be found in tabular form in a more conventional but possibly less clear manner in Tab. B.1 in Section B.1 in the appendix). The two single-height features "Middle" and "Max" were almost consistently outperformed by feature selection using multiple height values in all three metrics. Only LDA and kNN yielded F1-scores above 0.8 with LDA scoring the best F1-scores using the Steps features. The mean accuracy was between 0.61 and 0.95 for all approaches, but F1-score and G-mean varied to a larger degree (between 0.37 and 0.88 and between 0 and 0.78).

Fig. 7.8 shows the F1-score in detail. Multi-height approaches yielded better F1-scores results than single-height results. Additionally, kNNs, as an intuitive and explainable classification approach, could cope with most other classification approaches when used in conjunction with the Center-Steps. The easiest explainable

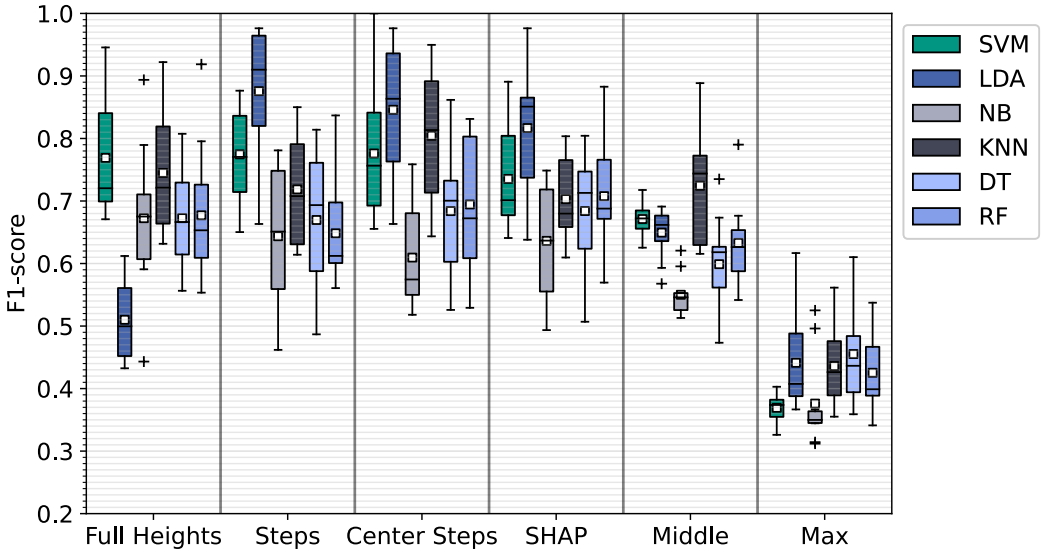


Figure 7.8: Boxplots for the classification approaches using F1-score. The two single-height classification methods score consistently lower performance values compared to the multi-height features.

classifiers DT, RF, and kNN showed similar performance despite their simple structure compared to more sophisticated approaches such as the SVMs. The standard deviations and general spread of all multi-height methods were larger than for the single height methods, showing that some runs achieved only low results. However, no methods scored F1-scores above 0.88.

7.4 Discussion

The measurement height showed an influence on the two cephalometric parameters CI and CVAI and consequently on the classification performance. A common frame of reference is therefore important to ensure comparability of CI and CVAI values among subjects. The naïve single-height classification features on the CI and CVAI values could not discriminate well between the four classes. A similar pattern could be observed if the control group was replaced with a left and right plagiocephaly class (see Chap B in the appendix) which even increased performance slightly. However, this cannot be compared quantitatively, because the dataset size was also reduced.

Overall, classification performance using only cephalometric parameters was poor, but using multiple height values instead of single-height cephalometric measurements increased classification performance for almost all classifiers. No feature selection clearly outperformed another. For head assessment with an automatic classification, multiple height measurements should be used which supports related studies on CT imaging [23, 146] including suggestions to add additional parame-

ters [150, 151]. Although the "Full-Heights" features technically contain the most information about the patient geometry (since the most height values are used), it only scored average classification results. The cause of this is likely a high redundancy in the measurements due to the close proximity.

The white box kNN classification performed similar to the competing approaches, indicating that in such a scenario, white box classification is indeed an attractive possibility. However, as it will be revealed during the course of this thesis, when comparing with more sophisticated approaches (such as the ones developed in Chapter 8 and Chapter 9 using a statistical shape model and a convolutional neural network), the cephalometric approach is outperformed. This is not surprising per se, but rather consistent with the observations of this study, namely that multiple height-based measurements improve classification performance as those classifiers similarly use a 3D representation of the patient.

7.5 Conclusion

In this work it was shown that when assessing cephalometric parameters and classifying different types of craniosynostosis using ML approaches, the extraction height of the CI and CVAI has an influence on the classification results. Using multiple extraction heights increased classification performance, especially if using the Steps, Center-Steps, or SHAP features and LDA. kNN-based classifiers showed a similar performance compared to non-explainable models on the same cephalometric parameters. Overall, F1-scores were comparatively low which suggests that more sophisticated methods are required for an adequate classification of craniosynostosis. While an explainable classification is important, a high classification performance is the most important prerequisite.

Chapter 8

Classification Using a Statistical Cranium Model

This chapter quotes partly in verbatim from the related open-access publication licensed under CC-BY in Diagnostics [140]. The original publication comprises both the publicly available model [141] and the statistical shape model (SSM)-based classification approach based on the model, originally published with a reduced dataset of 367 samples. For this thesis, they were separated into two different chapters, one for the SSM (Chapter 6) and one for the classification approach (Chapter 8).

8.1 Introduction

As outlined in the introduction of this thesis (see Section 1.1), the usage of data-driven machine learning (ML) models is a popular and versatile approach for diagnostic classification tasks. Additionally, as outlined in Chapter 6, statistical shape modeling is a popular method to combine geometric modeling and statistical analysis, which has been successfully applied for shape quantification and statistical analysis of craniosynostosis patients. However, a combination of the two modalities, i.e., an ML classifier making use of the extracted statistical information of the SSM, has not been proposed on 3D surface scans. The most similar approach used an SSM and handcrafted features with computed tomography (CT) data for classification [29].

The goal of this chapter is to develop an SSM-based classification approach for 3D surface scans. It builds on the proposed SSM (see Chapter 6) and fuses some ideas of the CT-based approach [29], such as the cranium segmentation, and the usage of the shape parameter vector with the proposed morphing pipeline (see Chapter 6).

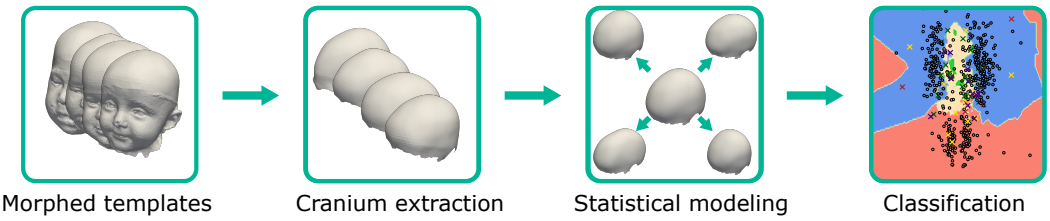


Figure 8.1: The craniosynostosis classification is based on a statistical model of the cranium. From the SSM in Chapter 6, the morphed templates were used, the craniums were extracted, the SSM was built and the classification could be performed.

8.2 Methods

Fig. 8.1 gives an overview of the pipeline from the morphed templates resulting from Chapter 6 to the creation of a cranium-only SSM and the subsequent classification of craniosynostosis.

8.2.1 Dataset and Preprocessing

For the description of the dataset, recordings, pathologies, class distribution, preprocessing steps, and landmarks, it is referred to Chapter 5. The dataset resulted in 496 samples of four classes: control, coronal suture fusion, metopic suture fusion, and sagittal suture fusion. The pipeline yielding the morphed templates is described in Chapter 6. The $N = 496$ samples from the original dataset, which had been morphed with the Laplace-Beltrami regularized projection (LBRP) morphing approach, were used for the classification approach in this chapter.

Intuitively, it is apparent that for the classification approach it is desirable to reduce the classifier input to only the relevant features. In this case, the cranium was required and the other parts of the scan such as face, ears, and neck had to be removed. The cranial part was defined manually on the reference head shape.[1] Points below the eyes and closer than $25\,\mathrm{mm}$ to ears and eyes were discarded. After extracting cranium points of all templates, an SSM of the cranium (a cranium model)

[1] As the morphed templates shared the same point identifiers due to correspondence establishment, once defined regions were applicable to all morphed templates.

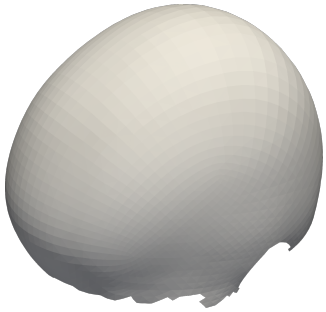


Figure 8.2: Mean shape of the SSM of the cranium. The ears have been removed to have the principal components focus on the cranium alone.

was computed.[2] This resulted in the number of points $p_{\mathrm{Cr}} = 2711$ for each morphed template. The mean shape of the cranium model is depicted in Fig. 8.2.

8.2.2 Statistical Modeling

The SSM was created on the extracted cranium to obtain the principal components of the SSM using weighted principal component analysis (WPCA) as described in the fundamentals (see Chapter 4) and the statistical model of the full head (see Chapter 6). From the observation matrix $\mathbf{S}_{\mathrm{Cr}} \in \mathbb{R}^{3p_{\mathrm{Cr}} \times N}$, the shape parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^{N}$ should be obtained which would be fed into the classifiers.

$$\mathbf{S}_{\mathrm{CrZM}} = \mathbf{s}_{\mathrm{Cr}} - \bar{\mathbf{s}}_{\mathrm{Cr}} \tag{8.1}$$

$\mathbf{S}_{\mathrm{CrZM}}$ denotes the zero mean observation matrix of the cranium models, $\mathbf{s}$ the shape, and $\bar{\mathbf{s}}_{\mathbf{Cr}}$ the mean shape.

$$\mathbf{s}_{\mathrm{Cr}} = \bar{\mathbf{s}}_{\mathbf{Cr}} + \mathbf{V}_{\mathrm{Cr}} \boldsymbol{\Lambda}_{\mathrm{Cr}}^{\frac{1}{2}} \boldsymbol{\alpha}_{\mathrm{Cr}} \tag{8.2}$$

$\mathbf{V}_{\mathrm{Cr}}$ the model's principal components, $\boldsymbol{\Lambda}_{\mathrm{Cr}}$ the eigenvalues of the sample covariance matrix, and $\boldsymbol{\alpha}_{\mathrm{Cr}}$ the shape parameter vector.

The first principal components of the statistical model (see Fig. 8.3) depicted the expected shape changes and pathologies: shape size, sagittal and metopic pathology, and asymmetric plagiocephaly shape deformities can be observed in the first three principal components. Compared to the SSM of the full head (see Chapter 6), the plagiocephaly shape differences were more pronounced. From the SSM, the shape parameter vector $\boldsymbol{\alpha}$ could be computed as

[2] Two inferior alternatives would have been possible to compute the cranium model: The first option would have been to extract the cranium before shape morphing, but this would be a difficult landmark-free morphing on a rather homogeneous surface, likely worsening the morphing results. The second option would have been to remove the facial points *after* computing the principal components, in this case the principal components would have been the same as in the head model (with reduced expressiveness on the cranium).

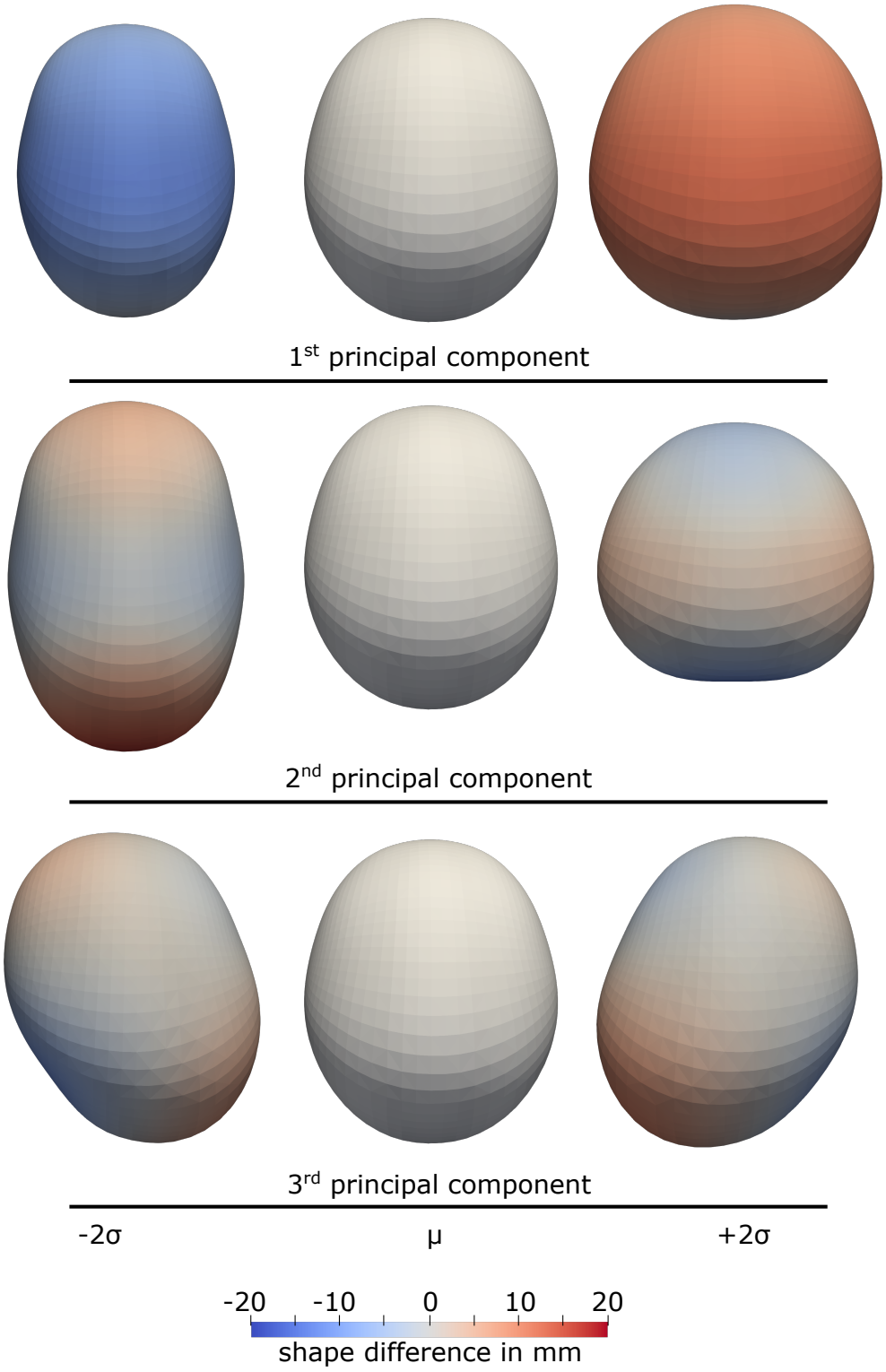


Figure 8.3: Principal components of the cranium SSM, the colorbar indicates the Euclidean differences to the mean shape. From left to right: -2σ, μ, and 2σ. From top to bottom: the first three principal components.

$$\boldsymbol{\alpha} = \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{V}^{-1} \left(\mathbf{s} - \bar{\mathbf{s}} \right), \tag{8.3}$$

and served as the input for the classification approaches.

8.2.3 Classification Setup

Using the coefficient vector as shape descriptors and as a direct input for a support vector machine (SVM) had been successfully tested in a different domain [152]. Since during the creation of the SSM, principal component analysis (PCA) was employed, the shape parameter vector $\boldsymbol{\alpha}$ had already been normalized assuming a Gaussian distribution. Normalization is generally beneficial for non-probability-based classifiers (i.e., SVM, k-nearest-neighbors (kNN), decision tree (DT), and random forest (RF)), which are known to be sensitive with respect to the scaling of the input features.

For the description of the classifiers, it is referred to Chapter 3 in the fundamentals. The used classifiers were SVM, linear discriminant analysis (LDA), naïve Bayes (NB), DT, RF, and kNN. All classifiers were implemented using the Python module `scikit-learn` [153]. For the SVMs, a linear kernel was chosen and non-binary classification was modeled using six one-versus-one binary classifiers. LDAs did not have tunable hyper-parameters and used a multivariate Gaussian distribution with a different mean, but the same covariance matrix for each class. Each prediction was assigned to the class whose mean was the closest in terms of the Mahalanobis distance taking into account the prior probability of each class. NB did not have tunable hyper-parameters and assumed conditional independence between input variables. As for LDA, a Gaussian model was used to distinguish between classes. kNN classification assigned the test sample according to the $k = 5$ closest neighbors in Euclidean space. For tie-breaking, the nearest neighbor among the tied classes was selected. DTs and RF used a hierarchical, tree-like structure with a maximum depth of six.

During PCA or WPCA, the principal components were ordered according to their variance, so the first principal components described the overall shape while the last components contained mostly noise. This is a process inherent to PCA with the goal to extract the directions of highest variance. The noise could arise from real noise (such as incorrect morphing, limited mesh resolution, or acquisition errors during scanning), or from small variation in the data which is perceived as noise from the PCA (such as less frequent geometric variation in the dataset or minor morphology changes influencing only a limited number of points). However, based on the assumption that geometric changes in the pathology have an impact on a large part of the geometry, it was expected that the parameters responsible for a good classification were concentrated in the first components. The optimal number of principal components was systematically increased by iterating over the first 100 principal components and using only the most principal components up to each

iteration.[3] For each run, three different metrics were used: accuracy, G-mean, and F1-score. While accuracy is the most intuitive metric for many people, it does not take into account the dataset imbalance. G-mean, on the other hand, is strongly influenced by a misclassification of minority classes, while the F1-score is influenced by both (see also Chapter 3.4).

The classification was performed using stratified 10-fold cross validation with a fixed seed for the random number generator and reproducible splits for all classifier.

8.2.4 Data-Augmented Classification Setup

In addition to the stratified 10-fold cross validation, the dataset was augmented with the mirrored samples from the SSM from Chapter 6. Instead of the 496 samples, the doubled amount, $496 \cdot 2 = 992$ samples were therefore available. This required the creation of a statistical cranium model composed of all 992 samples which was created with the same parameters as the original model. To enable a quantitative comparison of the metrics without data leakage, the following data subdivision was used: For each split, the test set was only composed of the original, non-mirrored samples. However, the training set was augmented with the mirrored training samples. This way, each of the samples from the original set was used only once for testing without the possibility of data leakage (i.e., a mirrored instance from the test set appearing in the training). A total of $2 \cdot 100 \cdot 6 \cdot 10 = 12000$ (2 dataset scenarios, 100 principal components, 6 classifiers, and 10 cross-validation splits) classification runs were performed in total.

8.3 Results

8.3.1 Principal Components Dependency

First, the results without data augmentation with a varying number of the principal components are presented. The most visible is a rapid increase of classification performance from 0 to 10 principal components across all three metrics when more shape information was revealed to the classifiers. For accuracy and F1-score of kNN and RF classification (see Fig. 8.4 and Fig. 8.6), a rapid decrease could be observed if more than 20 components were used. G-mean showed the most rapid decrease and the highest jitter (see Fig. 8.5). NB classification performed best according to F1-score (0.939), G-mean (0.926), and accuracy (0.966). RF classification and the DT performed worst.

[3] Successively increasing the number of principal components is a brute-force approach, but as PCA inherently limits the search space to the first components, it is a scalable approach even for large datasets.

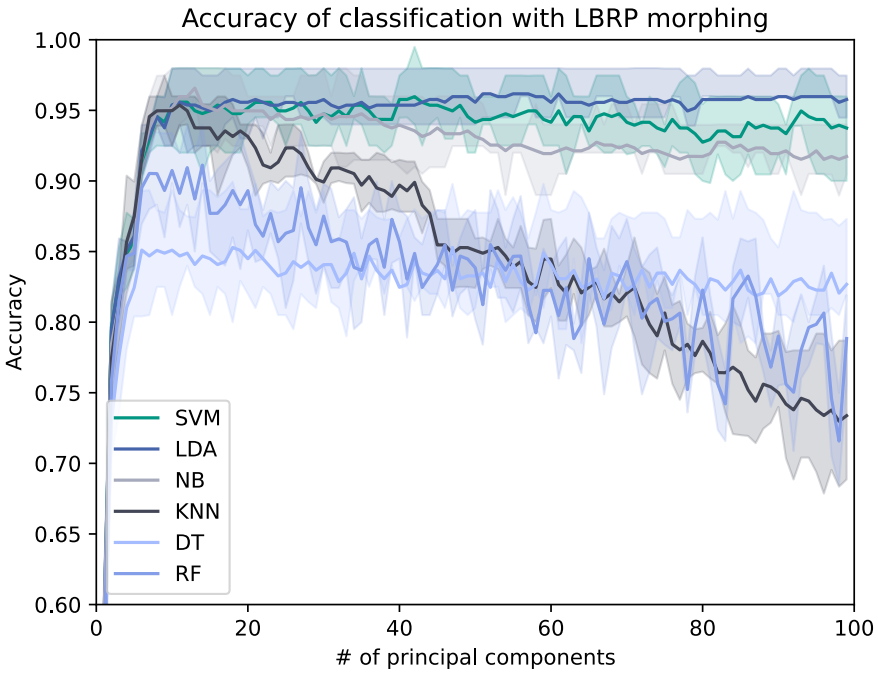


Figure 8.4: Accuracy as a function of the number of principal components used for the Laplace-Beltrami regularized projection (LBRP) classifier. Shown is the mean value and in lighter color the 25th and 75th percentiles.

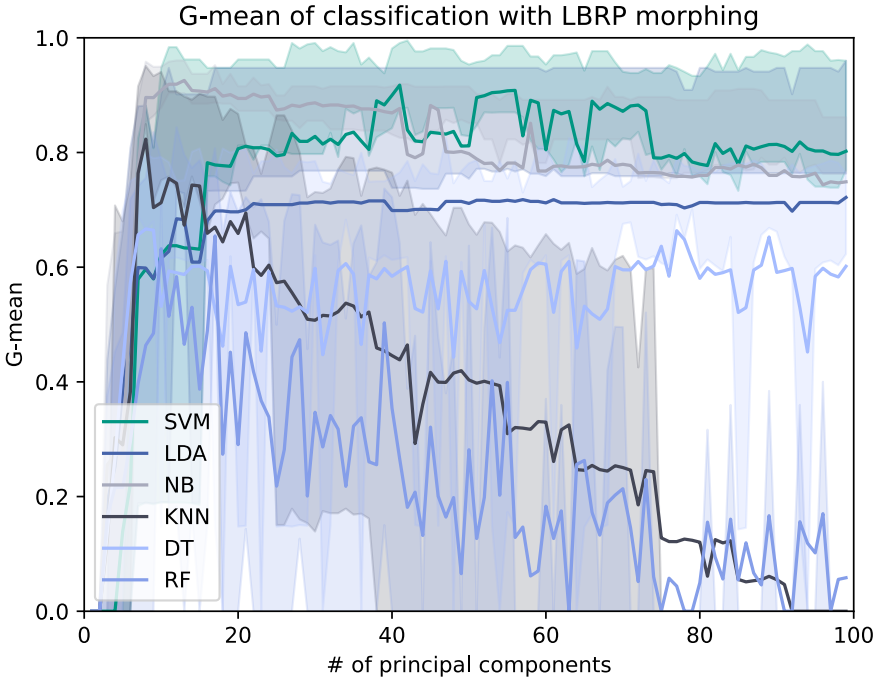


Figure 8.5: G-mean as a function of the number of principal components used for the Laplace-Beltrami regularized projection (LBRP) classifier. Shown is the mean value and in lighter color the 25th and 75th percentiles.

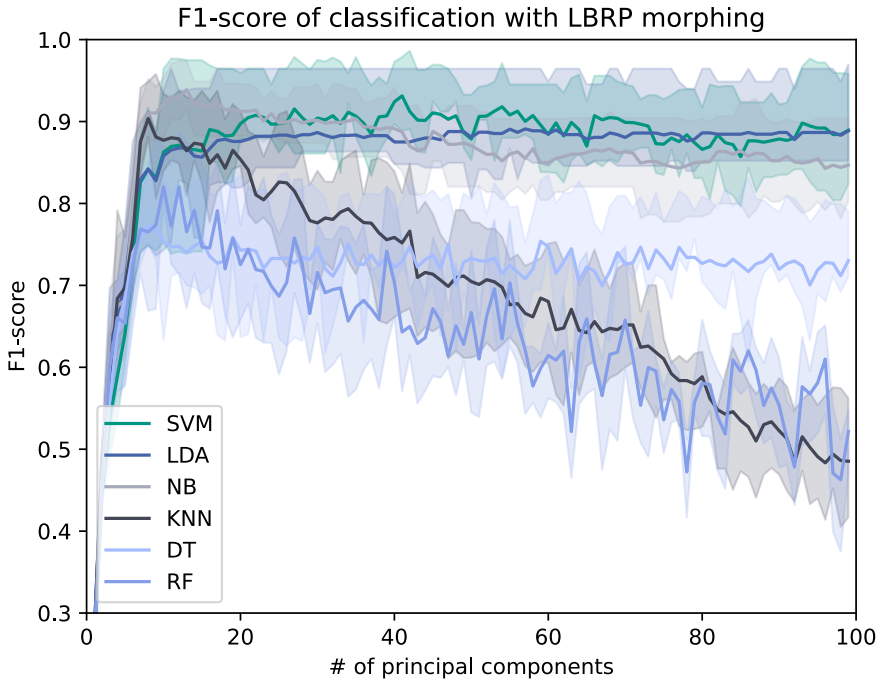


Figure 8.6: F1-score as a function of the number of principal components used for the Laplace-Beltrami regularized projection (LBRP) classifier. Shown is the mean value and in lighter color the 25th and 75th percentiles.

8.3.2 Data Augmentation

The results for the augmented dataset which includes the mirrored samples were consistently below the results on the original data for all classifiers on all metrics (see Tab. 8.1). Without data augmentation, the SVM, LDA, NB, and kNN classification yielded higher F1-scores than the results for the best multi-height approach (LDA with "Steps", see Chapter 7).

In the published initial classification study on the smaller dataset [140], LDA scored the highest accuracy (0.978, but on a smaller dataset, therefore not quantitatively comparable).

The same classification experiment was performed with the other morphing approaches and is available in the appendix in Chapter C. The model performance and error metrics showed slightly different values for classification performance, but the same trends could be observed: LDA, NB, and SVM classification performed best and classification performance decreased when increasing the number of components. Some results were even slightly higher, but the main trends were the same.

Table 8.1: Comparison of the classifiers on the cranium model. Displayed is cross validation mean ± standard deviation. On top are results without data augmentation and on the bottom are results including the mirrored dataset for evaluation.

Classifier	# components	Accuracy	G-mean	F1-score
Best classification run without patient mirroring				
SVM	n=40	0.958±0.029	0.918±0.086	0.931±0.058
LDA	n=56	0.962±0.032	0.718±0.365	0.891±0.101
NB	n=12	**0.966±0.022**	**0.926±0.053**	**0.939±0.049**
KNN	n=7	0.950±0.020	0.823±0.279	0.904±0.072
DT	n=7	0.851±0.047	0.667±0.235	0.773±0.076
RF	n=9	0.907±0.041	0.632±0.321	0.821±0.095
Best classification run with patient mirroring				
SVM	n=33	0.907±0.034	0.524±0.352	0.792±0.084
LDA	n=42	0.936±0.025	0.721±0.254	0.860±0.063
NB	n=13	0.938±0.031	0.858±0.083	0.897±0.066
KNN	n=8	0.809±0.055	0.522±0.273	0.703±0.119
DT	n=7	0.750±0.051	0.283±0.287	0.576±0.096
RF	n=5	0.798±0.031	0.049±0.147	0.575±0.038
Comparison with multi-height-classification (see Chapter 7)				
LDA	Steps	0.948±0.043	0.772±0.274	0.876±0.102

8.4 Discussion

This work proposed the first classification pipeline for craniosynostosis based on an SSM, tested on the largest dataset used for a study related to craniosynostosis to date. Multiple authors [30, 31] have shown that statistical shape modeling enables a quantitative analysis of the head shape with respect to craniosynostosis. In this work, it was demonstrated that SSM can not only quantify, but also classify head deformities. The classification results were better than multi-height approaches introduced in Chapter 7. Including the mirrored samples into the classification as data augmentation decreased classification performance.

Compared with the same study performed on an earlier version of the dataset with 367 subjects, the mean accuracy dropped slightly to 94% indicating that some dataset variation lead to slightly different results. However, the main conclusions of the original work remained the same [140], namely a good performance of the LDA and NB classifications and the inclusion of too many principal components lead to a performance reduction of the classifiers. As the classification approach of this thesis was tested with multiple morphing approaches and multiple classifiers on the largest currently available dataset, it was demonstrated that it is robust and does not rely on heavy hyper-parameters tuning.

Compared with other approaches from the literature, this approach performs similar, but slightly worse: Mendoza et al. [29] achieved a classification accuracy of 95.7% on 141 subjects using CT data and de Jong et al. [13] obtained an accuracy of

99.5 % on 196 samples using a feedforward neural network (FNN) in combination with ray casting and stereophotographs. As the classification approaches from the literature used different datasets, quantitative comparisons between different approaches are dataset dependent. For example, the datasets could systematically differ due to the usage of different scanning devices, admission protocols, or inclusion criteria of subjects favoring an "easier" or "more difficult" dataset.

One strength of an approach using an SSM is that they are a flexible and versatile modeling approach, which can aid clinicians in a variety of other tasks (such as patient counseling, visualization, or surgical guidance) as well. However, the control class of this study was assembled by the scans of children who visited the hospital without indication to be treated surgically. This included patients who were diagnosed with positional plagiocephaly. Thus, the control model represents a mixed group of children. The classifier might therefore have learned different decision functions as if the groups were separated.

Using PCA assumes that the training data follows a multivariate normal distribution. This assumption does not hold up for a head model which includes different pathology classes. With respect to the classification, PCA served as a reparameterization and ultimately as a dimensionality reduction procedure preferring less noisy input data in the first principal components. This seemed to be one of the key elements contributing to the success of this classification approach.

8.5 Conclusion

This work presented a craniosynostosis classification pipeline using the parameter vector of an SSM. State-of-the-art results comparable to both CT data and 3D surface scans were achieved and tested on the largest craniosynostosis-specific dataset to date. Morphing approaches showed little influence on the classification results, more important were the number of principal components and the classification approach. Data augmentation using the mirrored twins from the dataset lead to a performance drop. SSMs are flexible and versatile tools when working with clinical data and this approach expands the current use-cases of SSMs for classifying subclasses of head deformities. While using the shape parameters of a SSM for classification was only tested on head shapes for the classification of craniosynostosis, it is likely applicable to other shape families as well.

Chapter 9

CNN-Based Classification Using 2D Distance Maps

This chapter quotes partly in verbatim from the related open-access publication licensed under CC-BY in IEEE Transactions on Biomedical Engineering [154].

9.1 Introduction

In the introduction (see Chapter 1.2), it has been established that craniosynostosis is a condition affecting young infants and that the focus of this thesis is to improve radiation-free diagnosis of craniosynostosis. One already existing machine learning (ML) and successful approach for classifying craniosynostosis used a ray-based distance extraction scheme in combination with a feedforward neural network (FNN) [13], using a manually defined center point and achieving an accuracy of 99.5 % on an in-house dataset. It has to be noted that the values of metrics are dataset-dependent and cannot be compared quantitatively to other datasets. Furthermore, some disadvantages of classifying directly on the 3D data becomes apparent when considering data augmentation on 3D data. Data augmentation was limited to adding noise to the input features, while 3D transformations such as left-right patient mirroring and rotational misalignment had to be applied before training and cannot be randomized during training because the distance extraction is computationally expensive. The data augmentation is therefore not applicable dynamically during training.[1]

A popular choice for the classification of 2D images are convolutional neural networks (CNNs). They offer a flexible model design, and the easy adaption of many pre-trained models facilitating transfer learning. On head deformities, some

[1] Randomization before training leads to re-sampling of the generated instances during each epoch and is therefore inferior to dynamically employing randomization during training, in which all samples are truly random and different during each epoch.

CNN approaches include multiple viewpoints from different perspectives for craniosynostosis [38, 39] as well as a combination of segmentation and classification on plagiocephaly patients [143] with the goal of developing a mobile application. However, 2D photographs do not represent the full 3D head geometry. If CNNs were to be applied on the 3D geometry, a transformation from 3D to 2D is required, which had not been proposed for classifying head deformities. A suitable 2D image representation of the 3D head geometry could make use of the full head and combine 3D data with CNNs.

The contribution of this work is a mapping approach to obtain 2D images from 3D surface scans of the head which combined two ideas: asymmetry maps for plagiocephaly patients [34], and ray casting for distance extraction [13]. Those two ideas will be merged to create 2D distance maps and combined with a CNN classifier. Using 2D images instead of the original 3D geometry has some desirable properties when dealing with 3D patient data: patient anonymity is preserved (back-conversion would only yield a 3D scatter plot), and typical 2D image-based processing steps using filter kernels (such as interpolation, up-sampling with an under-sampled resolution, smoothing, or gradient computation) are enabled. As the 2D distance maps are subject to a defined coordinate frame, location-specific image processing can be applied, for example stronger smoothing in certain regions. The encoding of the 3D geometry into a 2D image enables using CNNs on the image domain for classification. Sophisticated network structures have been tested extensively on CNNs and there is a wide range of pre-trained networks available enabling transfer learning, which is usually considered helpful when dealing with small datasets. Image-based data augmentation strategies such as horizontal flipping, or horizontal shifting give more flexibility to the applicant. Data augmentation can be applied without much computational cost during training and enables additional randomization. 2D images can be re-scaled easily and it will be shown that classification performance can be maintained while systematically reducing image resolution. For the benefit of the community, the Python modules for the distance map creation were released, which can also be used on the previously published statistical shape model (SSM) [141].

This is the last chapter of this thesis concerned with introducing classification approaches and therefore compares the already employed classifiers of Chapter 7 and Chapter 8 as well as the FNN approach [13]. This enables a quantitative comparison of the current state-of-the-art classification approaches under the same conditions.

9.2 Methods

Figure 9.1 gives a full overview of the pipeline from the raw data to the distance map creation and the craniosynostosis classification. The dataset resulted in 496

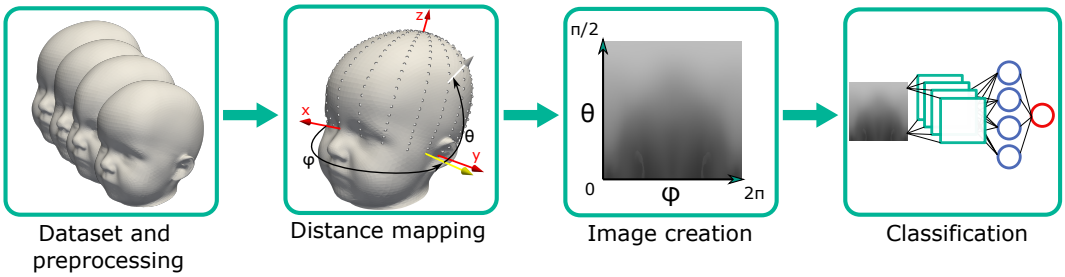


Figure 9.1: 2D distance map classification pipeline. Each dataset sample is preprocessed to remove corruptions. After distance extraction according to the mapping approach, the image can be assembled, and a CNN-based classification can be performed.

samples of four classes: control, coronal suture fusion, metopic suture fusion, and sagittal suture fusion. For the clinical acquisition of the dataset, the data distribution, and the preprocessing, it is referred to Chapter 5.

9.2.1 Distance Mapping

The patients' anatomic landmarks were used for the creation of a common coordinate system similar to the frontal, sagittal, and median planes. For a coordinate system, the sellion tragion orientation (STO) was chosen (see Section 5.4), as the sellions were located on different ends of the head and the midpoint between the two ears is approximately in the center of the head.

9.2.1.1 Spherical Mapping

The spherical mapping used a spherical coordinate transform for the ray creation to obtain the direction vectors $\mathbf{d_s}$

$$\mathbf{d}_\mathrm{s} = \left[\cos\varphi\cos\theta, \sin\varphi\cos\theta, \sin\theta\right]^\mathsf{T}. \tag{9.1}$$

The start point $\mathbf{p_s}$ of the ray was defined as the origin $\mathbf{p_c}$:

$$\mathbf{p}_\mathrm{s} = \mathbf{p}_\mathrm{c} \tag{9.2}$$

The two angle intervals were $0 \leq \varphi < 2\pi$ and $0 \leq \theta < \pi/2$ in the image domain. To retain the up-down relation of the distance map image, the image origin was placed in the bottom left corner and the direction for θ was defined from bottom to top (see Fig. 9.2, top right panel).

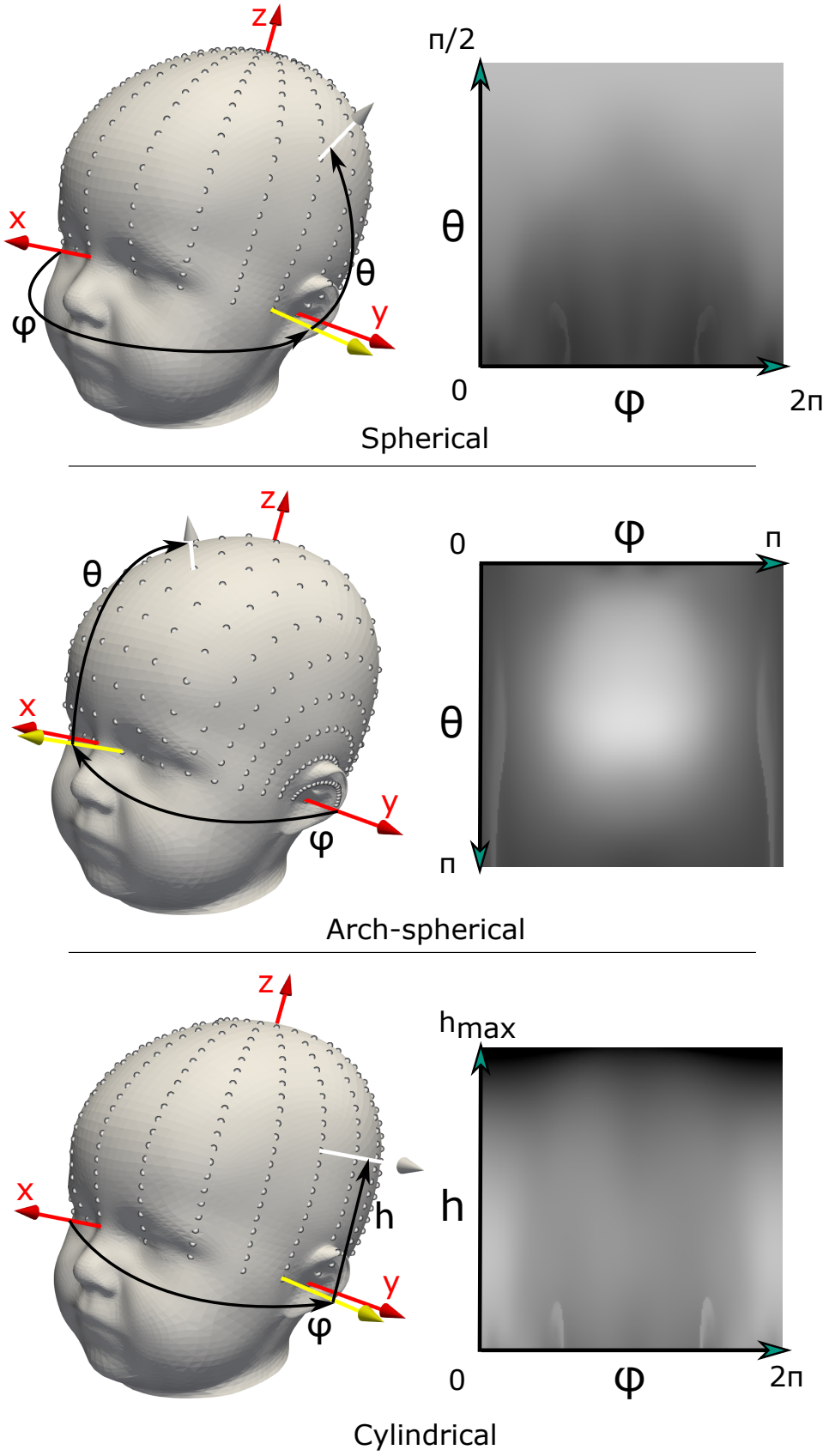


Figure 9.2: Visualization of the mapping types. Left: Angle definitions and coordinate frames. Hit points from the rays resulting from a 20×20 sampling are visualized. Right: Distance maps corresponding from the mapping with angle axes.

9.2.1.2 Arch-Spherical Mapping

The arch-spherical transform was designed to retain a frontal-vertical relationship when looking at the head from the top perspective and to provide a more regular sampling of the tip of the head with the direction $\mathbf{d_a}$

$$\mathbf{d}_\mathrm{a} = \left[\sin\varphi\cos\theta, \cos\varphi, \sin\varphi\sin\theta\right]^\mathsf{T}, \tag{9.3}$$

and the start point again being placed in the origin:

$$\mathbf{p}_\mathrm{s} = \mathbf{p}_\mathrm{c} \tag{9.4}$$

The corresponding angle intervals for φ and θ both ranged from $0 \leq \varphi < \pi$, but the φ direction was defined clockwise (mathematically negative) to retain the left-right relation. This way, the left part of the 2D image corresponded to the left part of the head when viewed from above and the back part of the head corresponded to the bottom of the 2D image.

9.2.1.3 Cylindrical Mapping

For the two spherical-based mappings, a grayscale gradient could be observed from the top to the bottom of the 2D image. By using a cylindrical-based mapping instead of a spherical one, the distance variation could be reduced for a larger part of the image. One key feature was that the center point was not constant, but moved toward the tip of the head h_max for each pixel row. Thus, the direction $\mathbf{d_c}$ was defined as

$$\mathbf{d}_\mathrm{c} = \left[\cos\varphi, \sin\varphi, 0\right]^\mathsf{T}, \tag{9.5}$$

and contrary to the previous approaches, the start point for each ray was defined as

$$\mathbf{p}_\mathrm{s} = \mathbf{p}_\mathrm{c} + h \cdot \mathbf{u}_\mathrm{z}. \tag{9.6}$$

The φ angle ranged from $0 \leq \varphi < 2\pi$ and h from 0 to the tip of the head h_max. Regardless of the mapping type, the angle intervals were sampled equidistantly. As with the spherical image, the reversed image direction (x-axis from bottom to top) of h was used to retain an up-down relationship.[2]

9.2.2 Image Creation

Each angle interval was sampled equidistantly in 224 steps, resulting in one ray direction for each of the 2D image pixels. The intersection of the 3D mesh surface

[2] A variant of the cylindrical approach with $h_\mathrm{max} = 1$ was also used in the cephalometric parameter extraction for the multi-height classification in Chapter 7.

with each ray was determind and the distance from the starting point to the hit point was extracted. Oriented bounding boxes trees from the `vtk` Python package [148] were used to speed up computation. The extracted distances were arranged in a 2D image grid, corresponding to the angle directions (e.g., φ and θ in the spherical mapping approach, see also Fig. 9.2). If multiple hit points were encountered (for example at the auricula, the outermost part of the ear), the minimal distance was chosen as the "correct" distance. If no hit point could be determined (for example due to corruptions in the scans), missing values were interpolated on the equally-spaced image grid as the mean of its four neighbors (which models missing pixels according to Laplace's equation) [155]. Small artifacts resulted from the tip of the head which were left unchanged. However, if required, they could be minimized on the image domain using smoothing. The actual image was created by converting the distances to integer pixel intensity values from 0 to 255. Two different normalization schemes were performed to transform the distance range to the required image intensity range: Linear re-scaling and per-pixel-based re-scaling.

Linear re-scaling used only *one* linear transformation for *all images and pixel* values. Mean distance and standard deviation were computed across all scans and distances (regardless of their ray orientation) to obtain one transformation to map the distances of $[-3\sigma, +3\sigma]$ to the image domain of $[0, 255]$. This way, the relationship between image intensity and distance is preserved, so short distances between center point and 3D surface correspond to low image intensities and, consequently, intensity gradients within the same image correspond to distance changes in the underlying 3D geometry.

The second approach, per-pixel re-scaling, used *one* linear transform for *each pixel position* (corresponding to one transformation per ray direction). This way, the relationship between image intensity and distance is *not* preserved, but the intensity range is better sampled for each pixel. The mapping is non-uniform, meaning that intensity gradients within one image do not correspond to distance changes in the 3D geometry, but instead correspond to intensity values relative to this pixel in the other images.

Fig. 9.3 shows the different scaling approaches for the same distance map for each pathology and Fig. 9.4 the different mapping types for one subject. Other normalization approaches might also be possible, e.g., scaling with respect to a control group or min-max scaling irrespective to other subjects. The Python source code[3] [156] for the distance map creation was made publicly available and can be combined with the previously published synthetic dataset [141].

[3]https://github.com/KIT-IBT/cd-map

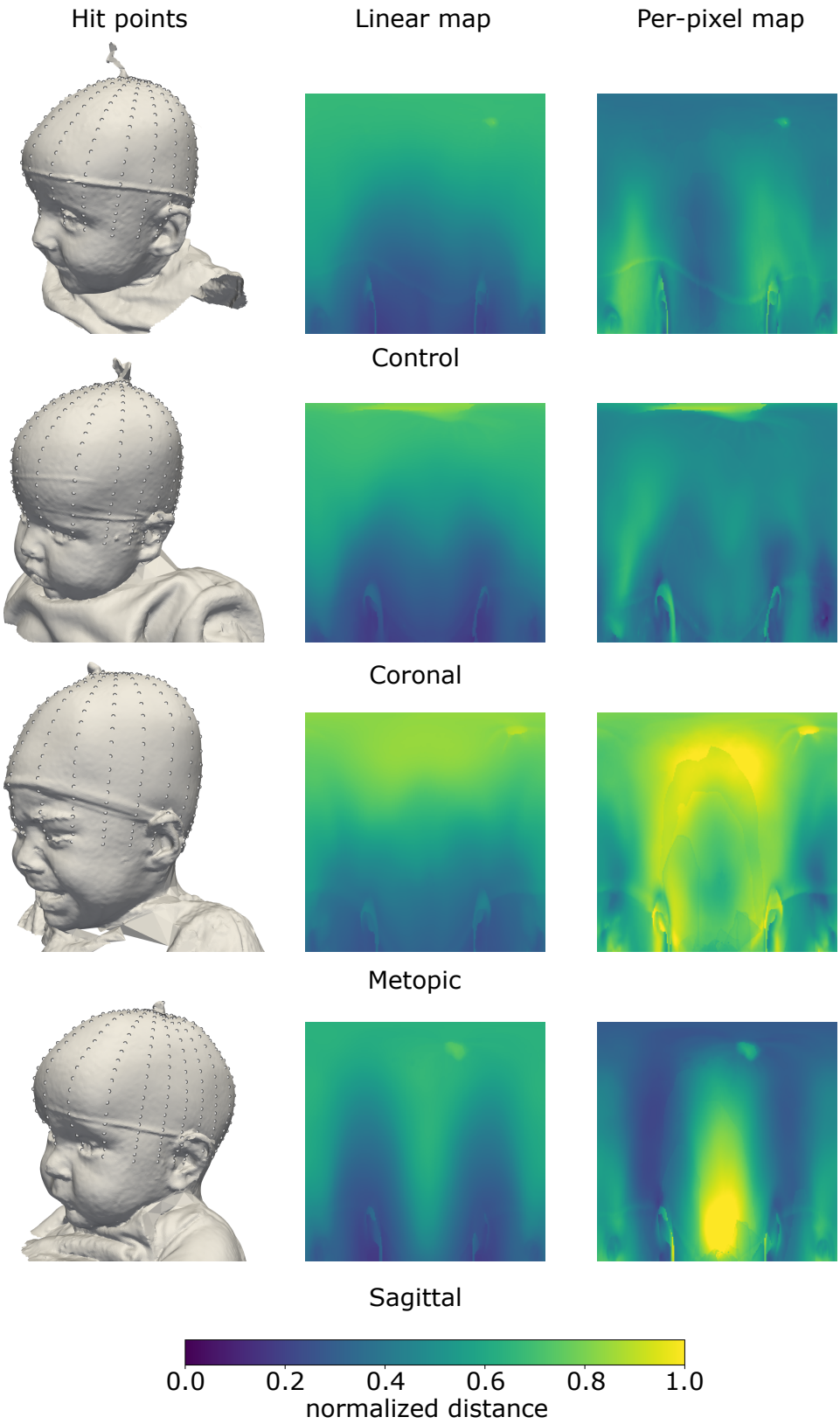


Figure 9.3: Linear and per-pixel scaling applied to the different pathologies using the spherical mapping. For the visualization of the hit points, 20×20 rays were used instead of 224×224. For visualization purposes, a blue-yellow colormap was used instead of grayscale.

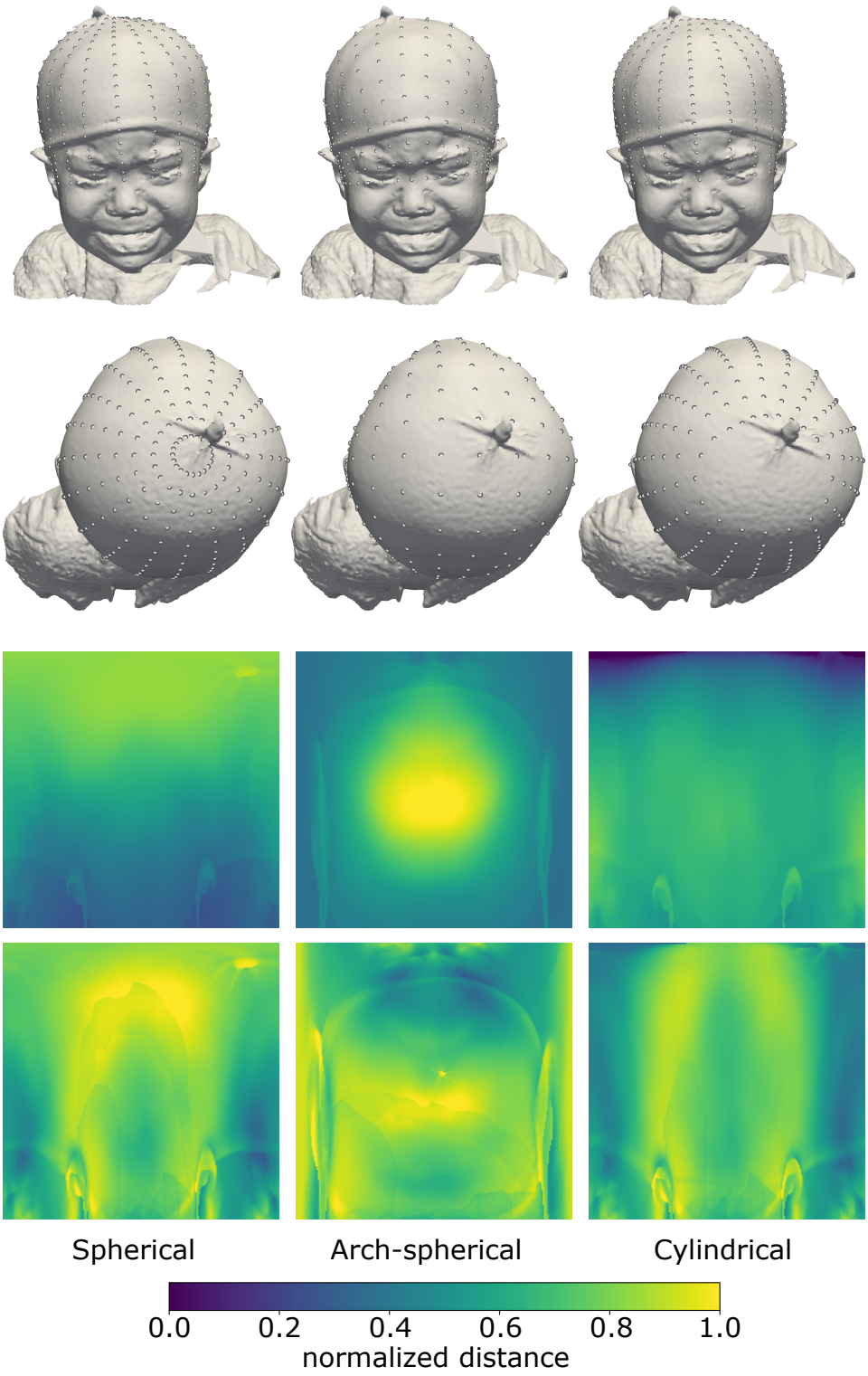


Figure 9.4: Mapping methods with different scalings for one subject. For the visualization of the hit points, 20×20 rays were used instead of 224×224. From left to right: Spherical, arch-spherical, and cylindrical. From top to bottom: Linear scaling and per-pixel scaling. For visualization purposes, a blue-yellow colormap was used instead of grayscale.

9.2.3 Experimental Setup and Network Training

The last step consisted of training a CNN on distance maps. Five test scenarios were considered and described in the next paragraphs: The first three evaluated classifier performance, while the forth and fifth scenarios studied the properties of the proposed method.

9.2.3.1 Classification Comparison

The first test was designed as a comparison between different CNNs using the proposed distance maps and alternative approaches from the literature.

A vanilla CNN trained from scratch only on the image data was considered. To find the optimal number of convolutional layers, it was trained with an increasing number of layers starting from 1 until 18. The highest metrics were scored by the CNN with 5 convolutional layers which is the one which was considered further. For the pre-trained CNNs, this included ResNet18 [157], AlexNet [94], GoogLeNet [158], and small and large MobileNet [159].

As alternative approaches, the white box SHAP-based multi-height approach using kNN classification (see Chapter 7) and SSM-based classification using naïve Bayes (NB) (see Chapter 8) and FNN-based classification [13] were considered. The FNN classifier [13] was re-implemented according to the original paper [13]. After initial testing, the dropout and batch-norm-layers were removed, which increased the performance metrics on the particular dataset used in this thesis. The FNN-based approach was tested on an icosphere-based extraction scheme as originally proposed [13] and on the distances extracted using the proposed mapping. This way, the FNN could be tested on both inputs.

9.2.3.2 Mapping Comparison

The three mappings and two scaling approaches were compared using the same network (pre-trained ResNet18) to test if the mapping had a substantial influence on the performance metrics. An exemplary set of images derived with the different mappings is shown in Fig. 9.4.

9.2.3.3 Data Augmentation Strategy

The image-based data augmentation was tested on pre-trained ResNet18, the vanilla CNN, and the FNN on the linear, spherical mapping. Four types of data augmentation were considered relevant: pixel noise, intensity noise, random flipping and random horizontal shift. Pixel noise was applied to each pixel as white Gaussian noise with standard deviation of $\sigma = 1/255$. Intensity-noise was applied to the full

image (making the image brighter or darker) as white Gaussian noise with $\sigma = 5/255$ and could be interpreted as enlarging or shrinking the full head. Random flipping in horizontal direction was applied with a probability of $p = 0.5$ and corresponded to a symmetric mirroring of the patients. The horizontal shift was designed as white Gaussian noise with standard deviation of $\sigma = 20/360 \cdot 2\pi$, shifting the image to one direction and inserting the cut-off part on the other side. This corresponded to a misaligned head rotation during recording, as if the subjects were looking slightly left or right.

Note that mirroring and shifting of 2D images could be performed "on the fly" during each training epoch. Before, mirroring and rotating on 3D data had to be performed before the training had started, could not be adjusted during training, and the data loaders were required to make sure that the mirrored samples stayed in the same training or test set. On 2D images, the created images were different during each epoch and the randomization could be adjusted during training, making it overall more flexible.

Fig. 9.5 visualizes the resulting data augmentation methods applied onto one subject and the resulting points back-transformed onto the 3D geometry.

9.2.3.4 Resolution

The resolution reduction comparison was designed to reduce computational cost for the distance map creation. As 224×224 is a standard size for CNN input images, the original approach used one ray per pixel. Using only $n \times n$ rays with n ranging from 7 to 224 in steps of 7 was tested. The smaller images were interpolated with intermediate points to obtain the CNN input dimension of 224. Three interpolation methods were tested: nearest-neighbor-mapping, bilinear and bicubic image interpolation (Fig. 9.6). Again, a pre-trained ResNet18 and the linearly scaled spherical mapping were used to ensure comparability among the experiments.

9.2.3.5 Attribution Maps

To make the model more interpretable, integrated gradients [160] were chosen as a visualization method to project the network's decision to the input image. Additionally, interpretability approaches might be able to rule out possible overfitting caused by a focus on unimportant parts of the head such as the ears (which are only expected to play a major role for coronal synostosis or plagiocephaly). The Python `captum` package [161] was used for computing integrated gradients. For visualizing the heatmap resulting from the integrated gradients on the 3D head surface, each point of the 3D surface was projected onto the image, was bilinearly interpolated, and the attribution value was back-projected to each 3D point. The following three different back-transformations for the spherical, arch-spherical, and cylindrical transformation yielded:

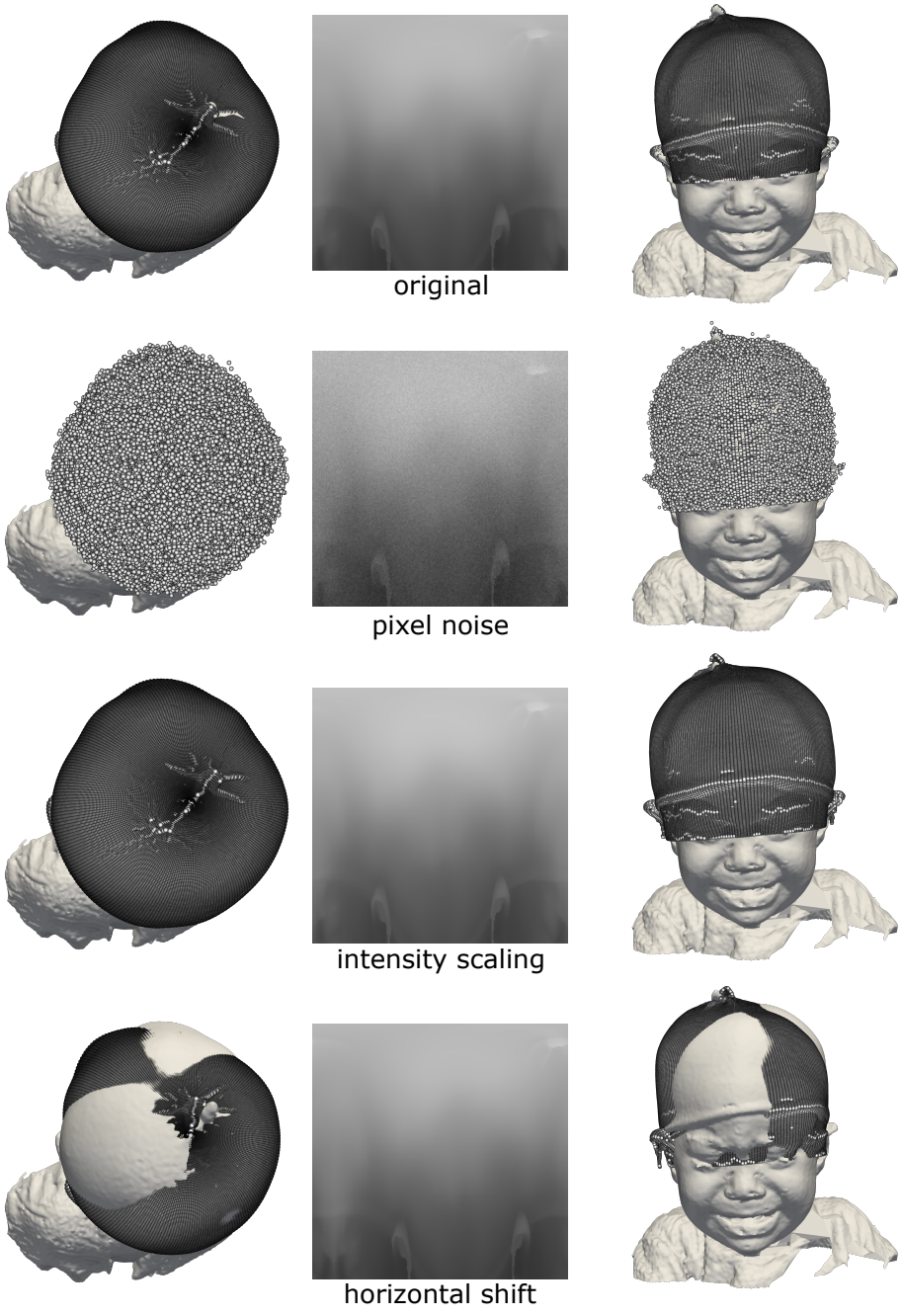


Figure 9.5: Data augmentation of the 2D images visualized and back-transformed into the 3D space, visualizing, how the extracted points from the 2D distance maps would appear as 3D scans. From top to bottom: Original, pixel noise (with standard deviation $\sigma = 5$), intensity noise making the head appear larger or smaller (with standard deviation $\sigma = 10$), horizontal shift ($19\,\mathrm{pixel} \approx 30\,°$) which translates into 3D as rotational noise.

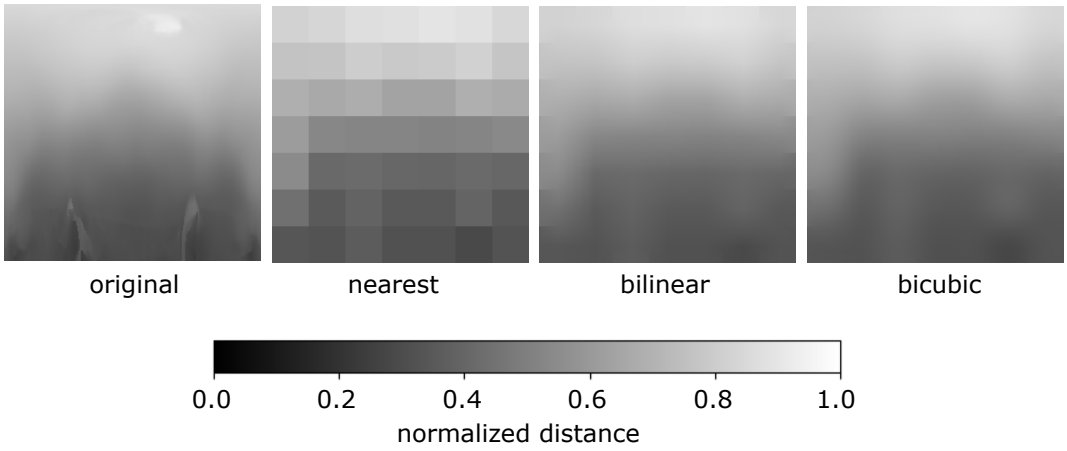


Figure 9.6: True 224×224 image in comparison with the three different interpolation methods nearest neighbors, bilinear, and bicubic from a 7×7 image.

$$\left[r, \varphi, \theta\right]^{\mathsf{T}}_{\text{spherical}} = \begin{bmatrix} \sqrt{x^2 + y^2 + z^2} \\ \arccos\left(z/r\right) \\ \operatorname{atan2}\left(y, x\right) \end{bmatrix} \tag{9.7}$$

$$\left[r, \varphi, \theta\right]^{\mathsf{T}}_{\text{arch}-\text{spherical}} = \begin{bmatrix} \sqrt{x^2 + y^2 + z^2} \\ \arcsin\left(y/r\right) \\ \operatorname{atan2}\left(z, x\right) \end{bmatrix} \tag{9.8}$$

$$\left[\varphi, \rho, z\right]^{\mathsf{T}}_{\text{cylindrical}} = \begin{bmatrix} \operatorname{atan2}\left(y, x\right) \\ \sqrt{x^2 + y^2} \\ z \end{bmatrix} \tag{9.9}$$

9.2.3.6 General Training Strategy

All classification scenarios were carried out using stratified 10-fold cross validation. The same random number generator was used for each experiment ensuring a consistent split of train and test samples across all different classifiers for the same fold. All networks were trained with cross entropy loss, adaptive moment estimation (ADAM) optimizer, a batch size of 32, and weight decay of 0.63. The initial learning rate for AlexNet was $1 \cdot 10^{-4}$ and for all other networks $1 \cdot 10^{-3}$. Pre-trained networks were trained with 300 epochs and a step size of 10, while from-scratch networks were trained with 1000 epochs and a step size of 100. The SSM-based classifier was trained with the same hyperparameters as in the previous work (see Chapter 8 or [140] on the new dataset). The python packages `pytorch` [89] and `scikit-learn` [153] were used for the implementation and `pytorch`'s pre-trained models had been trained on ImageNet [162]. Accuracy, G-mean, and macro F1-score were used for

model evaluation. Mean values and standard deviations were computed across all cross validation splits.

9.3 Results

9.3.1 Classification Comparison

As summarized in Table 9.1, all neural network (NN) classifiers showed good performance with mean accuracies above 0.95. CNN-based classifiers generally performed better than the competing approaches. The highest accuracies, G-means, and F1-scores were achieved by GoogLeNet and ResNet18. Standard deviations for F1-score computed across the ten folds were lowest for GoogLeNet and ResNet18 (indicating smaller disturbances for different training conditions) and increased for the other networks. The CNNs scored higher accuracies, G-means, and F1-scores than the FNNs. In general, accuracy ranged from 0.948 to 0.984 which corresponded to 18 fewer misclassified test samples. The confusion matrix with sensitivities and specificities of the pre-trained ResNet18 classifier is presented in Table 9.2.

Training times for each cross validation split measured on a high performance cluster running Red Hat Enterprise Linux using an Nvidia Tesla V100 were also included in Table 9.1. GoogLeNet required the longest training (306 s). In comparison, distance extraction for a 224×224 image took on average 102 s using a single thread on an Intel Xeon Gold 6230 processor. However, multiple scans could be processed in parallel since they were independent of each other.

9.3.2 Mapping Comparison

Classification results for different mapping approaches using the pre-trained ResNet18 are displayed in Table 9.3. All accuracies were 0.976 or above. All three metrics were consistently higher than the classification approaches in Table 9.1 except GoogLeNet. For the arch-spherical and cylindrical approach, the per-pixel mappings performed slightly better than the linear approach, but were in the range of one standard deviation.

9.3.3 Data Augmentation Strategy

Table 9.4 shows the classifier performance using “on the fly” 2D image-based data augmentation, compared to the networks without data augmentation (Table 9.1). All classifiers improved G-mean and F1-score. FNN showed the largest increase in all three metrics.

Table 9.1: Classifier comparison on linear, spherical mapping. For each metric, cross validation mean ± standard deviation is displayed.

Classifier	Training time	Accuracy	G-mean	F1-score
Alternative methods developed during this thesis (Chapter 7 and 8)				
Multi-height LDA-Steps	0.5 s	0.948±0.043	0.772±0.274	0.876±0.102
NB on SSM	5 s	0.966±0.022	0.926±0.053	0.939±0.049
Alternative methods from the literature				
FNN [13] on semi-icosphere	130 s	0.954±0.031	0.696±0.355	0.876±0.104
FNN [13] on distance maps	180 s	0.960±0.033	0.729±0.370	0.895±0.110
CNN-based methods using distance maps				
MobileNet small (pre-trained)	195 s	0.964±0.032	0.737±0.375	0.900±0.110
MobileNet large (pre-trained)	225 s	0.964±0.025	0.824±0.284	0.921±0.081
AlexNet (pre-trained)	241 s	0.970±0.016	0.829±0.287	0.923±0.076
Vanilla CNN	213 s	0.974±0.022	0.864±0.291	0.943±0.077
GoogLeNet (pre-trained)	306 s	0.982±0.014	0.938±0.078	0.962±0.042
ResNet18 (pre-trained)	210 s	**0.984**±0.020	**0.943**±0.070	**0.964**±0.043

Table 9.2: ResNet18 accumulated confusion matrix with linear, spherical mapping. For sensitivity and specificity, mean and standard deviations computed across all folds are depicted.

True class	Predicted class				Sensitivity	Specificity
	Control	Coronal	Metopic	Sagittal		
Control	276	0	1	1	0.993±0.014	0.977±0.031
Coronal	4	20	0	0	0.833±0.211	1.000±0.000
Metopic	0	0	70	0	1.000±0.000	0.995±0.014
Sagittal	1	0	1	122	0.984±0.033	0.997±0.008

Table 9.3: Mapping and scaling approaches using ResNet18. Displayed is cross validation mean ± standard deviation.

Mapping	Scaling	Accuracy	G-mean	F1-score
Spherical	Linear	0.984±0.020	0.943±0.070	0.964±0.043
	Per-pixel	0.976±0.012	0.944±0.035	0.962±0.018
Arch-spherical	Linear	0.976±0.018	0.938±0.078	0.959±0.044
	Per-pixel	**0.986**±0.018	0.959±0.069	0.974±0.041
Cylindrical	Linear	0.976±0.020	0.958±0.038	0.960±0.032
	Per-pixel	0.984±0.012	**0.971**±0.037	**0.978**±0.020

Table 9.4: Classifier comparison on linear, spherical mapping using image-based data augmentation. Displayed is cross validation mean ± standard deviation.. The second line for each classifier shows the increase compared to Table 9.1.

Classifier	Accuracy	G-mean	F1-score
FNN [13] on distance maps	0.968±0.024	0.888±0.083	0.928±0.055
(Mean improvement w.r.t. Table 9.1)	**(+0.008)**	**(+0.159)**	**(+0.033)**
Vanilla CNN	0.974±0.026	0.926±0.086	0.949±0.053
(Mean improvement w.r.t. Table 9.1)	(+0.000)	(+0.062)	(+0.006)
ResNet18 (pre-trained)	0.986±0.016	0.969±0.041	0.975±0.029
(Mean improvement w.r.t. Table 9.1)	(+0.002)	(+0.026)	(+0.011)

9.3.4 Resolution

In Fig. 9.7, the cross validation mean of accuracy, G-mean, and F1-score over pixel resolution are displayed for the ResNet18 classifier with spherical mapping and linear scaling. Starting with a pixel resolution of 14 and higher, G-mean was 0.81 or higher, accuracy 0.96 or higher, and F1 score 0.92 or higher. Using 14 steps in the interval of φ and θ instead of 224 steps in a 256-fold computation reduction of rays while bicubic image interpolation still yielded a G-mean larger than 0.95. All three interpolation methods jittered slightly and with similar amplitudes.

9.3.5 Attribution Maps

Fig. 9.8 visualizes the mean attribution across all scans for the different mappings. All three mappings assigned attribution to the frontal part of the head where typical deformations of sagittal and metopic craniosynostosis can be observed. The precise location varied on the mapping and was slightly shifted to the right for the spherical

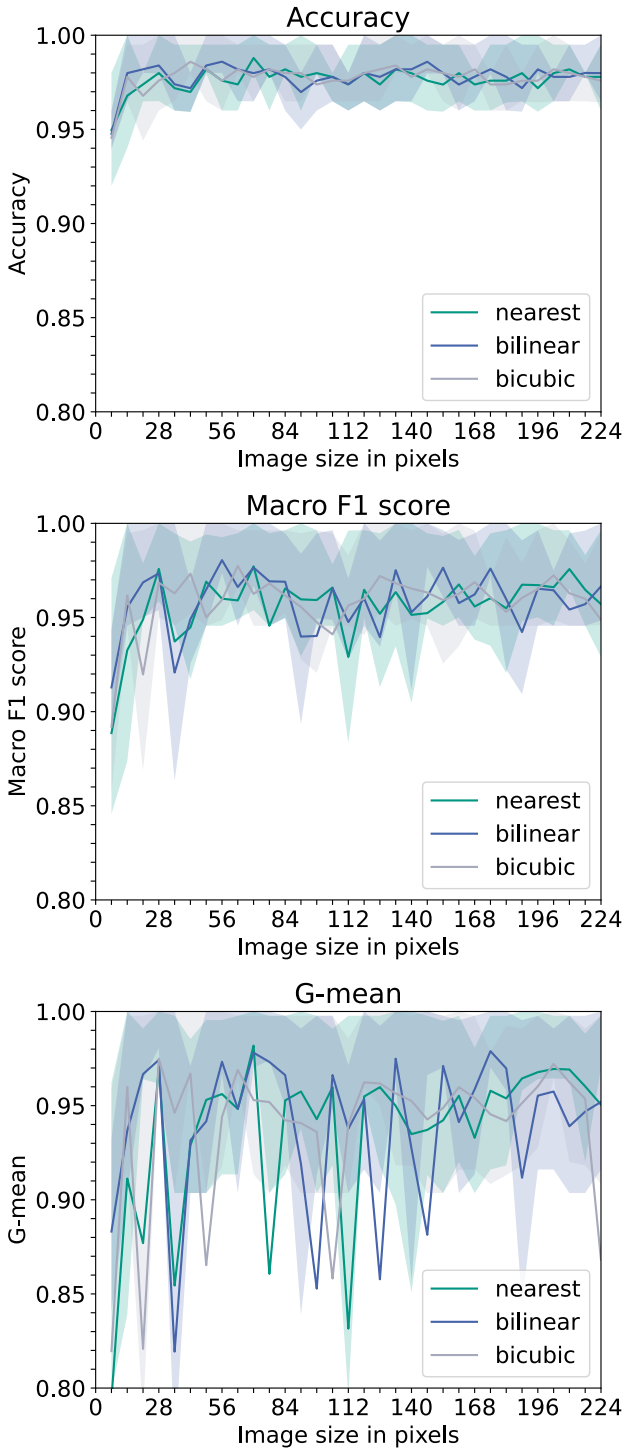


Figure 9.7: Mean cross validation metrics as functions of the number of pixels p to create a $p \times p$ image. Three different interpolation methods were used to create an up-scaled image: Nearest neighbors, bilinear, and bicubic interpolation.

and cylindrical mappings, and slightly shifted to the left for the arch-spherical mapping.

9.4 Discussion

A flexible mapping approach to create 2D distance maps from 3D head geometries was introduced, which was used for the classification of craniosynostosis. Multiple mapping variants with different coordinate systems and scaling approaches were proposed. While CNNs had been used for camera pictures from above [39], the proposed conversion encoded 3D data into intensity values on a 2D grid and enabled the first CNN for classifying craniosynostosis on the implicit 3D geometry. This was also used for the first systematic study to investigate the effects of reducing the 2D image resolution (and consequently sampling frequency of the 3D head surface) for classifying craniosynostosis. The 2D distance maps enable the usage of "on the fly" data augmentation methods typically employed for CNNs and can be used as an intermediate visualization before a machine learning classifier is employed, which preserves patient anonymity.

Using pre-trained networks was effective, especially ResNet18 showed good performance and scored highest in all three metrics. However, a vanilla CNN trained from scratch could outperform MobileNet and AlexNet and showed that pre-training is beneficial, but not a prerequisite for good classification performance. The network choice showed a larger influence on the three metrics than mapping choice (spherical, arch-spherical, or cylindrical) or scaling choice (linear or per-pixel scaling). This indicates that there is no "better" transformation, for the CNN, as long as the geometry is represented in the image. The mapping type might be more relevant when considering data augmentation methods, as only the spherical and cylindrical mapping allow a horizontal shift for rotation misalignment. Image-based data augmentation lead to an increase of G-mean and F1-score for all three tested classifiers. The original FNN classifier could be improved when using 2D image data and even more when introducing data augmentation during training. Taking into account the standard deviations, ResNet18 and GoogLeNet showed the most consistent performances across all ten folds while other classifiers showed higher standard deviations for G-mean and F1-score.

Using different image resolutions revealed that a low-resolution sampling of the head surface with a resolution of 14 steps per angle interval still showed classification results with G-mean and accuracy above 0.95 for bicubic up-scaling. This corresponds to a 256-fold ray-reduction of triangular ray intersection. Classification could be performed with substantially lower resolution than previously performed on 3D surface scans. Since low-resolution images represent low spatial frequencies well and suppress high spatial frequencies, it suggests that low spatial frequencies

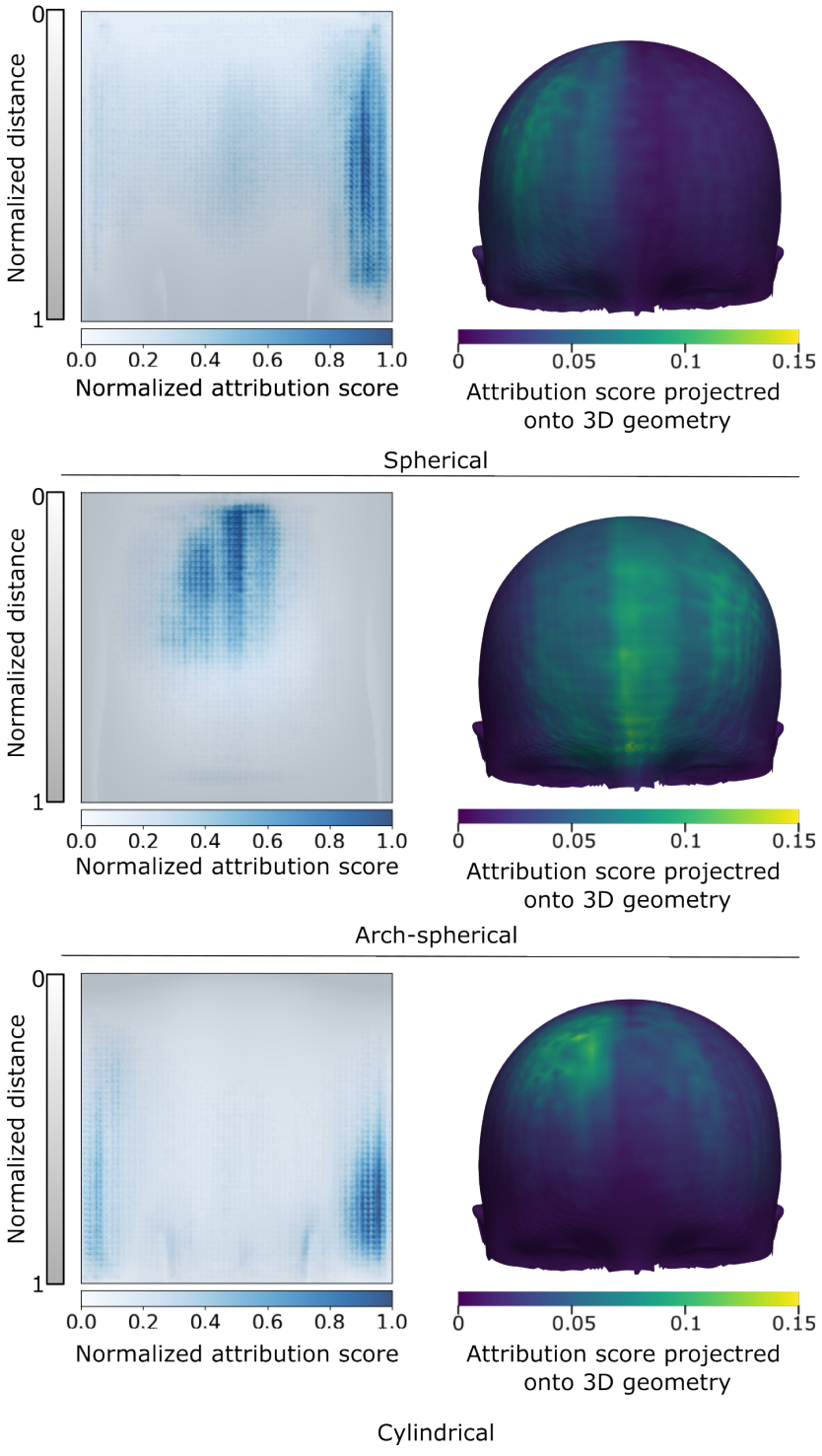


Figure 9.8: Mean attribution for all subjects in the image domain with a transparent overlay of the map (left) and projected onto the 3D surface for the respective mapping (right). From top to bottom: Spherical, arch-spherical, and cylindrical mapping. The 2D image shows a transparent overlay of the grayscaled distance map as a visual guide, while attribution is colored in blue. A larger value means higher attribution.

are most relevant for the classifiers. High-resolution artifacts (which may result from the ears or the often visible tip resulting from the caps) might be weakened. The reduction of input parameters for machine-learning-based classifiers might be a promising follow-up study and pave the way toward an interpretable classifier trained on few, carefully selected features. Although the resolution study was performed only on the spherical linear mapping, the results are likely valid for the other mappings as they showed little influence on the classification performance overall. The observed accuracy fluctuation of 1–1.5 % for the different resolutions was likely caused by different network training conditions, although all samples among splits were kept consistent. Low-resolution images might reduce the required precision of scanning devices or enable domain transfer to computed tomography (CT) imaging, even with high slice thickness.

One reason for the success of the CNNs might be that the filter kernels on the 2D image ensure that the classifier is trained on locally confined features. This impedes the simple correlation of spatially not connected input pixels and might be beneficial for classification performance. In contrast, FNNs interpret the image as a large 1D feature vector, thus allowing the creation of features based on random correlations across the image. A second reason might be that pre-trained networks might cope more easily with the small amount of data often present in medical classification problems. Especially ResNet18 and GoogLeNet seemed to be able to effectively fine-tune the fully connected layers after pre-training. However, the vanilla CNN proved to be an effective classifier without using pre-training and even surpassed some of the pre-trained network architectures.

Attribution maps intend to provide insights of how the classifier made its decision and suggest that the CNN was indeed triggered by features specific to the condition. Qualitatively, parts of the head which would be considered less important by humans (such as the ears) were assigned only little attribution. Higher attributions were assigned to the forehead with a prominent spot on either the left or right side of the head, corresponding to pathological differences between the classes, suggesting that the network makes use of geometric relevant parts of the head. It has to be noted that attribution mapping is generally not a replacement for explainable classification and generalizations from attribution mapping need to be interpreted carefully [163].

There are also limitations to this study. As with many studies in biomedical engineering, the used dataset contains only some hundred samples, even though it is the largest dataset of craniosynostosis patients used in a classification study to date. Optimally, multiple datasets should be used to further validate the models, which might increase trust in parents and physicians. However, as craniosynostosis head scans show the face of the patients, there are currently no publicly available clinical datasets and data sharing is complicated due to patient data regulations. Other groups might make use of the publicly available SSM [141]. Data augmentation or data synthesis might be an option to make the classification models as robust as

possible. It was shown that image-based random horizontal flipping and a random horizontal shift during training increased classification performance for CNNs and FNNs alike.

The three proposed distance map variants sample the 3D geometry with equidistant angle intervals which leads to non-equidistant sampling intervals on the 3D surface, resulting in more points at the tip of the head (see Fig. 9.4). However, this apparently did not hamper classification performance. One disadvantage of the 2D distance mapping is the reliance on the three manually annotated landmarks. Future work should focus on automatic registration of the scans, for example using random sample consensus (RANSAC).

In general, this mapping approach is not tied to 3D surface scans and might be used in other domains, for example for CT scans, head shape analysis or any shape analysis for spherical-like objects. Distance maps from magnetic resonance imaging (MRI) and CT could be used for classification purposes or combined with surface scans to obtain a cross-domain dataset from all three modalities. Especially the domain transfer to CT scans seems promising: It seems likely that low-resolution-maps from existing CT or x-ray imaging can achieve similar results, potentially reducing ionizing radiation if a radiography is still desired or inevitable.

9.5 Conclusion

Distance mapping approaches to transform 3D head shape information to 2D intensity-encoded images were presented which were combined with CNN-based classifiers for craniosynostosis. The conversion to 2D images enables the usage of "on the fly" data augmentation (horizontal mirroring and shifting) and pre-trained CNNs, and preserves patient anonymity. Resolution of the images was reduced systematically which showed that using the 2D image structure, low-resolution images could be used for classification without a substantial decrease of classification accuracy. ResNet18 achieved the best classification performance, showing that 3D surface scans are suitable for a reliable classification of the most common types of craniosynostosis. Although this mapping encoded 3D surface scans, it is not inherently confined to this domain and could be used for a combined image-based classification dataset. To facilitate this process, the Python source code was published as free and open source software.

PART IV

IMPACT OF DATA SYNTHESIS STRATEGIES

Chapter 10

Classification of Craniosynostosis Using Synthetic Training Data

This chapter quotes partly in verbatim from the submitted publication "Impact of Data Synthesis Strategies for the Classification of Craniosynostosis" which is currently under review.

10.1 Introduction

As stated in the introduction (see Chapter 1) and the clinical fundamentals (see Chapter 2), craniosynostosis is a rare disease and is included in the list of rare diseases by the American National Organization for Rare Disorders. Due to the low prevalence, strict patient data regulations, and difficulties in anonymization (3D surface recordings show head and face), there are no publicly available clinical datasets of craniosynostosis patients available online. Synthetic data could potentially be used as a substitute to develop algorithms and approaches for the assessment of craniosynostosis, but only one synthetic dataset based on a statistical shape model (SSM) developed during this thesis [141] has been made publicly available. Scarce training data and high class imbalance due to the different prevalences of the different types of craniosynostosis [12] call for the usage of synthetic data to support or even replace clinical datasets as the primary resource for deep-learning-based assessment and classification. The inclusion of synthetic data could facilitate training due to the reduction of class imbalance and increase the classifier's robustness and performance. Additionally, synthetic data may also be used as a cost-effective way to acquire the required training material for classification models without manually labeling and exporting a lot of clinical data. Using synthetic data for classification studies in a supporting manner or as a full replacement for clinical data has gained attraction in several fields of biomedical engineering (e.g. [164, 165]), especially if clinical data is not abundant. While classification approaches of craniosynostosis on computed

tomography (CT) data [29], 2D images [143], and 3D surface scans [13, 140, 154] have been proposed, the dataset sizes were below 500 samples and contained a high class imbalance. Using synthetic data is a straightforward way to increase training size and stratify class distribution.

However, although the need for synthetic data had been acknowledged [13], synthetic data generation for the classification of head deformities has not been systematically explored yet. In this work, it is aimed to test the effectiveness of multiple data synthesis methods both individually and as multi-modal approaches for the classification of craniosynostosis. As described in the fundamentals in Chapter 4 and during the development of the SSM during this thesis in Chapter 6, SSMs have been used successfully for quantification and classification of craniosynostosis and are a straightforward way to synthesize data. Although their value in the clinical assessment of craniosynostosis has been shown, the impact of SSM-based data augmentation for the classification of craniosynostosis has not been evaluated yet. With the introduction of a conversion of the 3D head geometry into a 2D image, image-based convolutional neural network (CNN)-based classification [154] can be applied on low-resolution images. This enables image generation using principal component analysis (PCA) or generative adversarial networks (GANs) [99]. GANs have been suggested as a data augmentation tool [13] and have been able to increase classification performance for small datasets [166].

The goal of this work is to employ a classifier based on synthetic data, using three different types of data synthesis strategies: SSM, GAN, and image-based PCA. The three modalities are systematically compared regarding their capability in the classification of craniosynostosis when trained only on synthetic data. It will be demonstrated that the classification of craniosynostosis is possible with a multi-modal synthetic dataset performing similarly to a classifier trained on clinical data. Additionally, a GAN design is proposed, tailored toward the creation of low-resolution images for the classification of craniosynostosis. Both the GAN, the different SSMs, and PCA, were made publicly available along all the 2D images from the synthetic training, validation and test sets.

10.2 Methods

10.2.1 Dataset and Preprocessing

The resulting data consisted of 496 samples of four classes: control, coronal suture fusion, metopic suture fusion, and sagittal suture fusion. Information about the dataset such as the data distribution and head shapes of each class are described in Chapter 5. In Chapter 9, a 2D encoding of the 3D head shape (“distance maps”) was

defined which was also included in the pre-processing pipeline with the spherical and linear variant of Chapter 5.

10.2.2 Data Subdivision

Three data generators were defined (GAN, SSM, and PCA) to synthesize synthetic data, on which the classifier should be trained. Only half of the clinical dataset (the validation set, see Fig. 10.1) was used as training data for the data generation models.

If the test set had also been included in the synthetic data sources, this would have lead to leakage (an overestimation of the model performance due to statistical information "leaking" into the test set). Instead, the schematic displayed in Fig. 10.1 was used, which comprised a stratified 50–50 split of the clinical data and used one half of the samples as the validation set and the other half as the test set.

The test set was separated from the validation set, only to be used for the final evaluation of the classifier. Following this approach, the test set did neither have any influence on the synthetic data, nor was it incorporated in the validation set and should therefore be a true representation of unknown data to the classifier. The validation set was used to select the best network during training and for hyperparameter tuning, but not as training material. Additionally it was used as the original (training) data which the synthetic image generators were built on. The synthetic training set was created from the validation set according to the three data synthesis approaches described below: SSM, GAN, and PCA. The three approaches operated on different domains (also depicted in Fig. 10.1): While the SSM was applied directly on the 3D surface scans, the GAN and the PCA used the 2D distance map images. All images were created as 28×28-sized craniosynostosis distance maps which was found to be sufficient for a good classification in the resolution study in Section 9.3.4. Each of the three individual approaches SSM, GAN, and PCA are described below.

10.2.3 Data Synthesis

10.2.3.1 Statistical Shape Model

To create the SSM data, the submodels of Chapter 6 were used. The pipeline for the SSM creation consisted of initial alignment, dense correspondence establishment, and statistical modeling to extract the mean shape and the principal components from the sample covariance matrix (see also Chapter 6 for more information).

For each submodel, the coefficient vectors were cut off after 95 % of the normalized variance to remove noise which ensured that only the most important

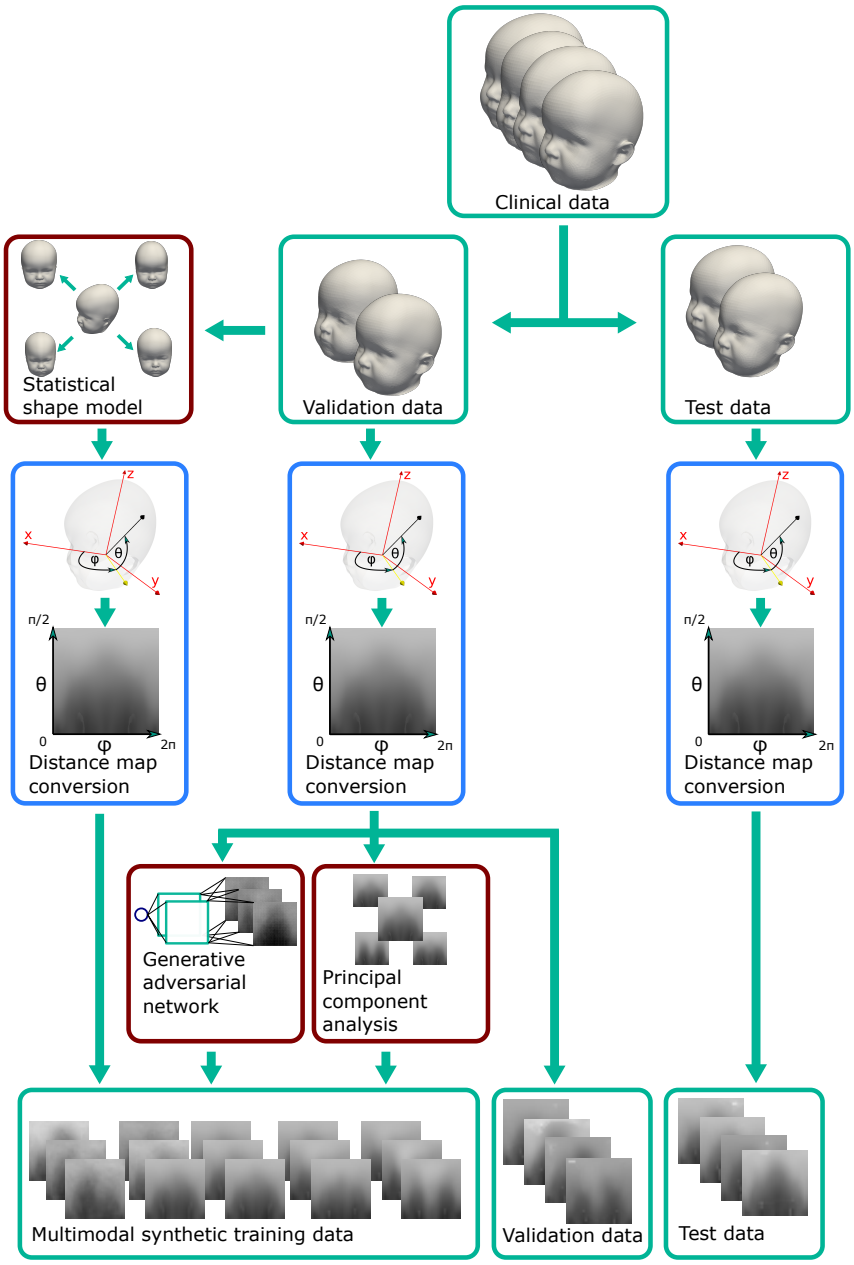


Figure 10.1: Data subdivision for the synthetic-data-based classification and the creation of synthetic data. The test set was separated initially from the dataset, while the validation set was used to produce the synthetic samples on which the CNN was trained. Green: data, blue: 3D-2D image conversion, dark red: generative models.

components were included in the SSMs. The synthesis of the model instances $\mathbf{s}$ could be performed as

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{V}_\mathrm{s}\boldsymbol{\Lambda}_\mathrm{s}^{\frac{1}{2}}\boldsymbol{\alpha}_\mathrm{s}, \tag{10.1}$$

with $\bar{\mathbf{s}}$ denoting the mean shape, $\mathbf{V}_\mathrm{s}$ the principal components, $\boldsymbol{\Lambda}_\mathrm{s}$ the sample covariance matrix, and $\boldsymbol{\alpha}_\mathrm{s}$ the shape coefficient vector. 1000 random shapes were created of each class using a Gaussian distribution of the shape coefficient vector and for each sample, a craniosynostosis distance map was created.

10.2.3.2 Image-Based Principal Component Analysis

Ordinary PCA was used as another modality to generate 2D image data. While the SSM also made use of PCA in the 3D domain, image-based PCA operated directly on the 2D images. This was a computationally inexpensive and less sophisticated alternative to both GANs and SSMs since neither extensive model training, nor hyperparameter tuning, nor 3D morphing, nor correspondence establishment was required. Ordinary PCA was employed for each of the four classes separately and was performed as

$$\mathbf{i} = \bar{\mathbf{i}} + \mathbf{V}_\mathrm{i}\boldsymbol{\Lambda}_\mathrm{i}^{\frac{1}{2}}\boldsymbol{\alpha}_\mathrm{i}, \tag{10.2}$$

with $\bar{\mathbf{i}}$ denoting the mean image in vectorized shape, $\mathbf{V}_\mathrm{i}$ again the principal components, $\boldsymbol{\Lambda}_\mathrm{i}$ the sample covariance matrix, and $\boldsymbol{\alpha}_\mathrm{i}$ the coefficient vector of the principal components. Again, the principal components were cut off after 95 % of the variance and 1000 random images were created by drawing from a Gaussian distribution.

10.2.3.3 Generative Adversarial Network

The GAN was designed as a conditional [105] deep convolutional [100] Wasserstein [106] GAN with gradient penalty [107]. It was trained for 1000 epochs using the Wasserstein distance [106] which is considered to stabilize training [167]. Instead of the originally proposed weight clipping, a gradient penalty [107] of $\lambda = 1$ was used. 10 critic iterations were employed before updating the generator and a learning rate of $3 \cdot 10^{-5}$ was used for both networks. The loss L could be described as follows [107]:

$$L = \mathbb{E}_{\tilde{x}\sim p_z} D(\tilde{x}|y) - \mathbb{E}_{x\sim p_r} D(x|y) + \lambda(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2 \tag{10.3}$$

with x denoting the real samples, $\tilde{x}$ denoting the generator samples $G(z|y)$, and $\hat{x} = \epsilon x + (1-\epsilon)\tilde{x}$ with ϵ denoting a uniformly distributed random variable between 0 and 1 [107]. The design in terms of the intermediate image sizes is visualized in Fig. 10.2. The full design including all layers is described in the appendix in Chapter D.2.

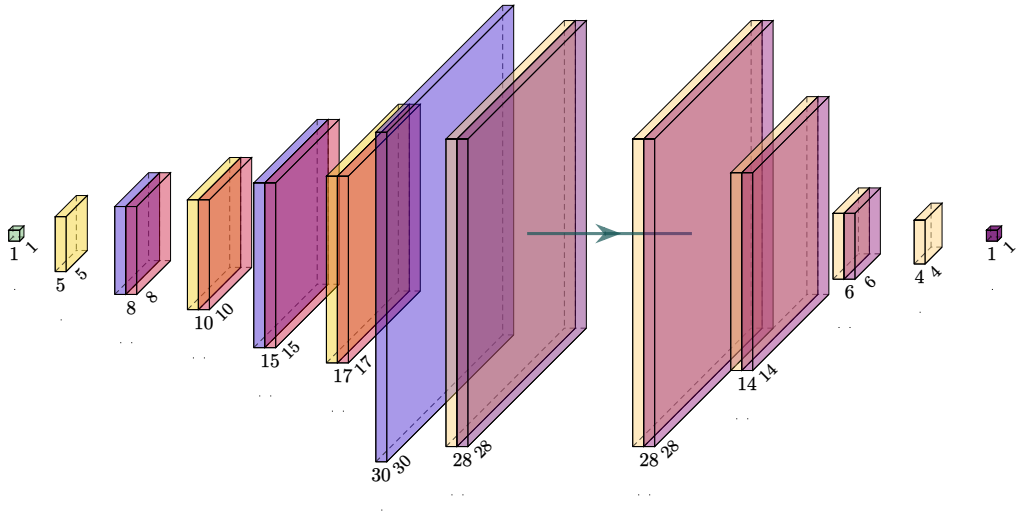


Figure 10.2: Visualization of the intermediate image sizes from the used GAN model. Left: generator, right: critic (discriminator). The filter kernel sizes are described in the appendix in Chapter D.2.

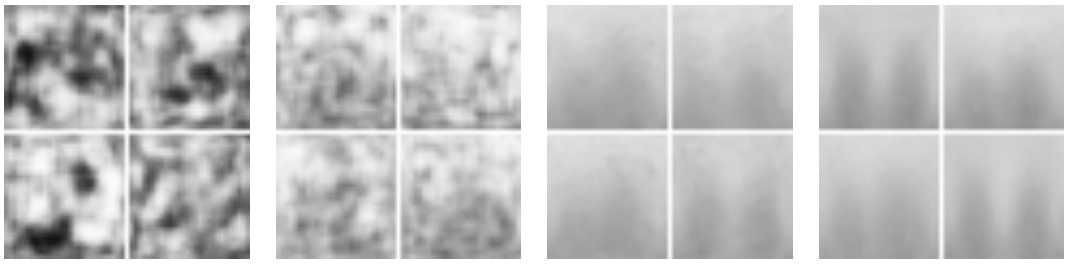


Figure 10.3: Progression of random images created from the GAN generator during different stages of training visualized as a 2×2 grid.

The network design was crafted using a mixture of transposed, interpolation, and normal convolutional filter kernels, aimed to prevent checkerboard artifacts and large patches. During the initial model design, checkerboard artifacts were observed using only transposed convolutions (present in a previous publication [168]), and large patches were observed using only upscaling layers (interpolations). The combination of interpolation layers and transposed convolutional layers lead to better images than each of the approaches alone (see Fig. 10.3 for the training image and Fig. D.1 in the appendix for the artifacts in other GAN images). The conditioning of the GAN was implemented as an embedding vector controlling the image label that was supposed to be synthesized.

10.2.4 Image Assessment

As a metric to assess the similarity of the synthetic images to the clinical images, structural similarity index measure to closest clinical sample ($SSIM_{cc}$) was defined as the structural similarity index measure for each *synthetic* sample by using the minimum $SSIM_{cc}$ with respect to each *clinical* sample of the same class N:

$$\mathrm{SSIM}_{\mathrm{cc},i} = \min_{\forall n \in N} \mathrm{SSIM}(p_{\mathrm{synthetic},i}, p_{\mathrm{clinical},n}) \tag{10.4}$$

It has to be noted that the $SSIM_{cc}$ itself did not assess the quality of the synthetic images, but was rather designed to evaluate the similarity to the clinical images. This approach had the goal to quantify a "good" data generator: The synthesized data should on the one hand not be "too" similar to the original data (because otherwise simply the original data could be used), but on the other hand, not too different either (because otherwise they might not be a true representation of the underlying class anymore). "Good" images should not be "too close" to 1, and not "too low".

10.2.5 CNN Training

All training data used in this study was made publicly available: The synthetic and clinical samples of this study were made available in its distance maps representation on Zenodo.[1] [169] The synthetic data generators (GAN, SSM, and PCA) were made available as `pickle` and `pytorch` files.[2] A Python script was included to create synthetic samples for all three image modalities, enabling users to create a large number of samples.

Resnet18 was used as a classifier since it showed the best performance on this type of distance maps [154]. The publicly available pre-trained Resnet18 model by `pytorch` [89] was used and the weights were fine-tuned during training. All images were bilinearly reshaped to a size of 224×224 to match the input size of Resnet18. Different runs of CNN training were performed on all seven combinations (three times one data source, three combinations of two data sources, and one combination of all data sources) of the synthetic data. The CNN was trained only on synthetic data (except for the clinical scenario which was trained on clinical data for comparison). During training, the model was evaluated on both the (purely synthetic) training data and the (clinical) validation set (see also Fig. 10.4). The best-performing network during training was chosen according to the maximum F1-score on the validation set. The test set was never touched during training and only evaluated in a final run after training.

[1] https://zenodo.org/record/8117499
[2] https://github.com/KIT-IBT/craniosource-gan-pca-ssm

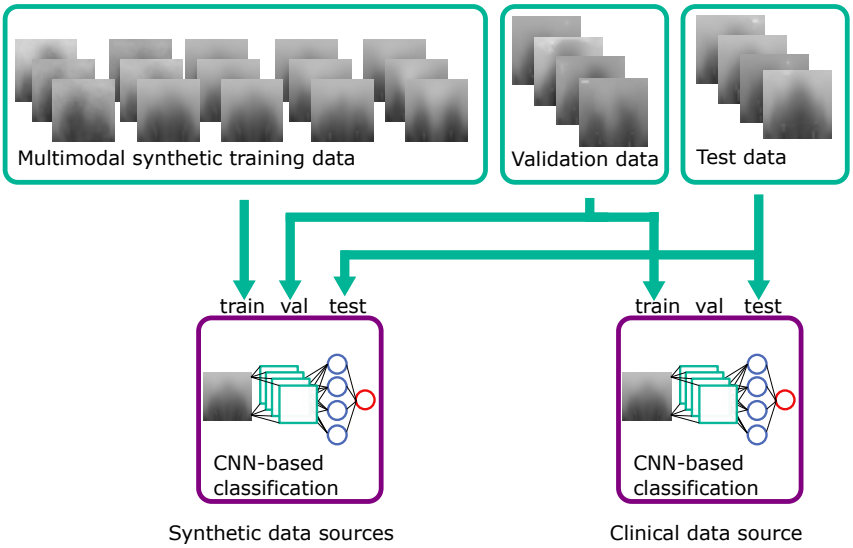


Figure 10.4: Classifier training using the synthetic data, the validation data, and the test set. The CNN classifier using clinical data uses the validation data as a training set. Green: data, violet: classification models.

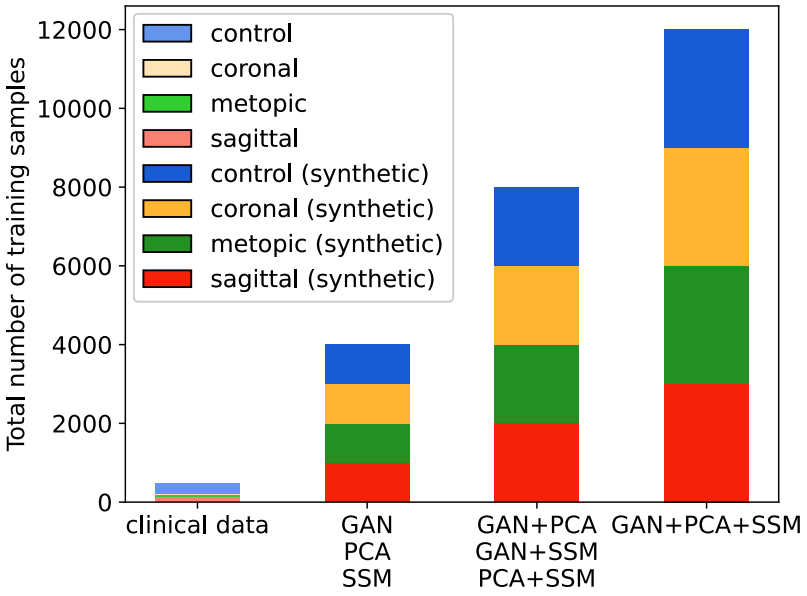


Figure 10.5: Number of training samples in each classification scenario. The color for the synthetic samples are darker. The clinical scenario has less than 500 samples while all synthetic scenarios have 4000, 8000, or 12000 samples.

As multiple data sources were used, the models had a different number of training samples (see Fig. 10.5) and all synthetically-trained models were trained for 50 epochs. The adaptive moment estimation (ADAM) optimizer, cross entropy loss, a batch size of 32 with a learning rate of $1 \cdot 10^{-4}$, and weight decay of 0.63 after each 5 epochs was used as hyperparameters. To evaluate the synthetically-trained models against a clinically trained model, one additional CNN was trained on the clinical validation data with the same parameters except a higher learning rate of $1 \cdot 10^{-3}$.

The following types of data augmentation were used during training: Adding random pixel noise (with $\sigma = 1/255$), adding a random intensity (with $\sigma = 5/255$) across all pixels, horizontal flipping, and shifting images left or right (with $\sigma = 12.44$ pixels). All those types of data augmentation corresponded to real-world patient and scanning modifications: Pixel noise corresponded to scanning and resolution errors, adding a constant intensity was equal to a re-scaling of the patient's head, horizontal flipping corresponded to the patients as if they were mirrored in real life, and shifting the image horizontally modeled an alignment error in which the patients effectively turned their head 20° left or right during recording.

10.3 Results

10.3.1 Image Evaluation

Fig. 10.6 shows images of each of the different data synthesis types compared with the clinical images. From a qualitative, visual examination, the synthetic images had similar color gradients, shapes, and intensities as the clinical images. GAN images appeared slightly noisier than the other images and did not show the left and right ear visible in the other images.

From the quantitative comparison (see Fig. 10.7), ordinary PCA images were substantially and consistently more similar to the clinical images than the other two modalities (differences of the medians larger than 0.02), while SSM and GAN images were characterized by lower SSIM_{cc} values with the SSM and the GAN being comparable except for the coronal class in which the SSM created the most dissimilar samples.

10.3.2 Classification Results

All presented runs were carried out on the untouched test set. Convergence was achieved already during the first ten epochs, indicating that there was sufficient training material for each model. According to the classification results for the synthetic training in Tab. 10.1, the SSM was the best single source of synthetic data

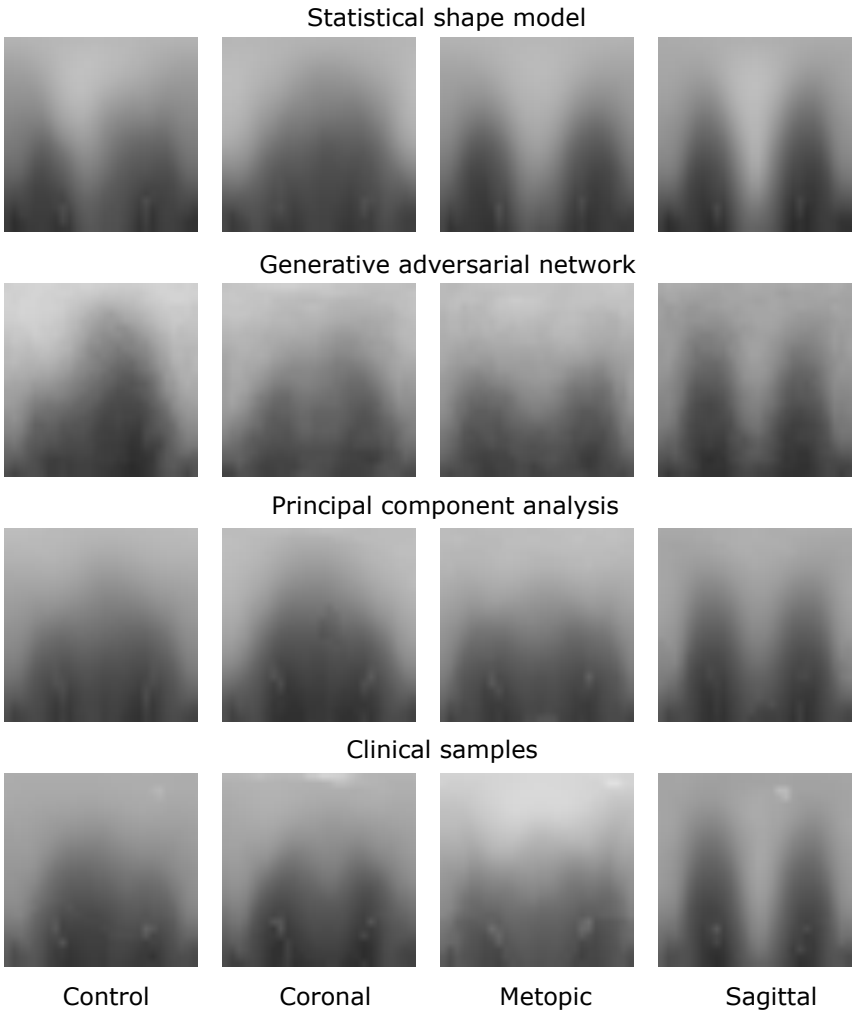


Figure 10.6: Images of all three data modalities and clinical samples. From top to bottom the image modalities: SSM, GAN, PCA, clinical. From left to right the four classes: Control, coronal, metopic, sagittal.

with an F1-score higher than 0.85. All combinations of synthetic models showed F1-scores higher than 0.78. The classifier on the clinical data scored an accuracy above 0.96, but was surpassed by the combination of GAN and SSM. F1-score was highest for the clinical classification with a value of 0.9533, but the combination of SSM and GAN scored a slightly lower F1-score of 0.9518. Including a second data source always increased the F1-score compared to a model with a single data source. The combination of SSM and GAN even scored a higher G-mean than the clinical data.

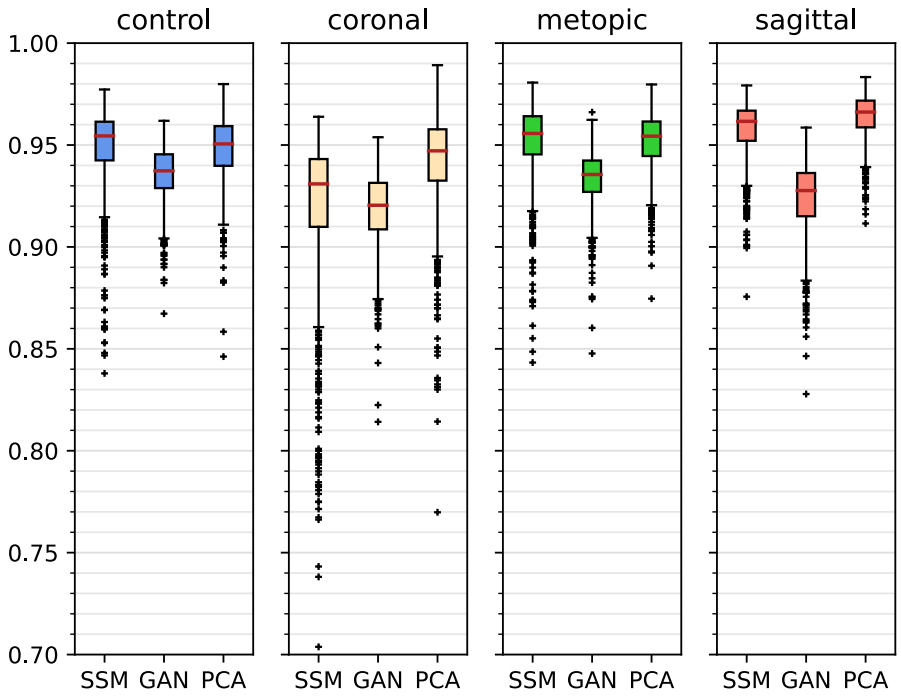


Figure 10.7: Boxplots of $SSIM_{cc}$ of each class for each of the synthetic data generators.

Table 10.1: CNN-classification comparison on the test set trained on different synthetic data sources. Boldface: best results among the synthetic data sources.

Synthetic data source	Accuracy	G-mean	F1-score
GAN	0.4274	0.0000	0.4930
PCA	0.7581	0.7910	0.6997
SSM	0.9153	0.9004	0.8547
GAN-PCA	0.8508	0.8543	0.7823
GAN-SSM	**0.9677**	**0.9609**	**0.9518**
PCA-SSM	0.9153	0.9125	0.8595
GAN-PCA-SSM	0.9597	0.9552	0.9445
Clinical	0.9637	0.9481	0.9533

10.4 Discussion

Without being trained on a single clinical sample, the CNN trained from the combination of the SSM and the GAN was able to correctly classify 95 % of the data. Classification performance with training on the synthetic data proved to be equal to or even slightly better than training on the clinical data, at least for the data generated using the SSM and the GAN (and optionally also including PCA). This suggests that certain combinations of synthetic data might be indeed sufficient for a classification algorithm to distinguish between types of craniosynostosis. Compared with the classification results from Chapter 8 and Chapter 9, the purely synthetic-data-based classification performs in a similar range and sometimes even better than other approaches on clinical data [13, 29, 39, 140, 154].

The SSM appeared to be the data source contributing the most to the improvement of the classifier: Not only did it score highest among the unique data sources, but it was also present in the highest scoring classification approaches. One reason for this might be that, according to the $SSIM_{cc}$, it was also the least similar data source for most of the classes. Due to the inherent modeling of the geometric shape in 3D, the created 2D distance maps were always created from 3D samples, while PCA and the GAN could, in theory, create 2D images which did not correspond to a morphologically correct 3D shape. In contrast, the GAN-based classifiers only showed a good classification performance when combined with a different data modality and its synthesized images seemed to show less pronounced visual features than the other two modalities. However, the $SSIM_{cc}$-based metric showed no substantial difference between the GAN images and the other two modalities. However, since the GAN training included images from all classes and the image label was determined by an embedding, features from different classes might appear in images from other classes. The PCA images were neither required, nor detrimental for a good classification performance. According to the $SSIM_{cc}$, the PCA images were the most similar images to their clinical counterparts.

Overall, a combination of different data modalities seemed to be the key element for achieving a good classification performance. Both SSM and PCA model data according to a Gaussian distribution, while the GAN uses an unrestricted distribution model. The different properties of modeling the underlying statistical distribution of a Gaussian distribution (SSMs and PCA) on the one hand, and without an assumed distribution (GAN) on the other hand might have lead to a symbiotic effect and compensation of their respective disadvantages increasing overall performance for the combination.

One limitation of this study is the small dataset. Using a smaller test set makes the results more prone to “lucky” dataset distribution. Cross-validation is costly and would have required training a differnt GAN and conducting the seven data source scenario during each of the cross-validation splits, which would have made

this study even more computationally expensive. Another piece of criticism can be made to the clinical classification scenario. As the clinical classification uses the same dataset for training and validation, this might make it prone to over-fitting. However, the resulting classification metrics achieved in this study were similar to a classification study on clinical data alone [154] which suggests that over-fitting has not been an issue. Another limitation is that the data from the SSM, GAN, and PCA were not synthesized dynamically during training. Since for each modality, the images were pre-generated, this limits the overall randomness. However, the data augmentation was dynamic, so the samples during each epoch were different.

10.5 Conclusion

It was shown that it is possible to train a classifier for different types of craniosynostosis based solely on artificial data synthesized by an SSM, a PCA model, and a GAN. Without having seen any clinical samples, a CNN was able to classify four types of head deformities with an F1-score higher than 0.95 and performed comparable to a classifier trained on clinical data. The key component in achieving good classification results was using multiple, but different data generation models. Overall, the SSM was the data source contributing most to the classification performance. For the GAN, using a small image size and alternating between transposed convolutions and interpolations were identified as key elements for suitable image generation. The datasets and generators were made publicly available along with this work. Clinical data are not required for the training of craniosynostosis classifiers, paving the way into cost-effective usage of synthetic data for automated diagnosis systems.

PART V

FINAL REMARKS

Chapter 11

Outlook

Automated clinical assessment of craniosynostosis and other head deformities will likely rely on multiple tools, for example automatic diagnosis using a convolutional neural network (CNN) and a visualization in clinical practice using a statistical shape model (SSM). Some specific steps toward further clinical applicability resulting from this thesis are outlined in this chapter. First, a path toward clinical applicability in the diagnosis of head deformities is presented, and second, a path toward related engineering challenges building on some of the developed methods during this thesis is presented.

The following list comprises some of the most important milestones for clinical applicability:

- **Multi-label classification:**
  Multi-suture synostosis and syndromic conditions should be included into the classifier. This could be implemented with a multi-label approach using an ensemble network with binary one-vs.-rest classifiers specialized to distinguish one pathology from everything else.
- **Distinction between healthy subjects and plagiocephaly patients:**
  Healthy subjects without plagiocephaly should become a large part of the database. Routine recordings of cleft lip or cleft palate patients might be a suitable and readily available data source.
- **Clinical prototype testing:**
  A prototype for usage during clinical admission should be constructed and tested. The recording of the patient could directly be classified and combined with an SSM for parent counseling showing the strongest deformations.

The following list compiles directions away from a classification problem, but still related to head deformities and to the research conducted during this work:

- **Time-dependent statistical modeling:**
  Multiple recordings of the same subject across multiple points of time could be used for a time-dependent statistical model. The model could predict head

growth after skull remodeling (for craniosynostosis patients) or during helmet therapy (for plagiocephaly patients). Related topics for SSMs are motion modeling for magnetic resonance imaging (MRI) tracking [170] and attribution modeling [110].

- **Skull inference of craniosynostosis patients using statistical coupling of shape and skull:**
  A common SSM of skull and head surface could be created on computed tomography (CT) scans. Using a posterior shape modeling [108] approach, the skull could be predicted from the 3D surface scan to estimate surgical incisions for therapy planning. A coupled skull-head model has been proposed [171], but was not applied to craniosynostosis.

Chapter 12

Conclusion

Three different classification approaches for craniosynostosis have been developed during this thesis. While the cephalometric multi-height approach combines clinically established parameters with simplicity and explainability using k-nearest-neighbors (kNN) classification, the approach failed to achieve the high classification performance of the other classifiers. This shows that the simple translation of established parameters to machine learning (ML) approaches is not enough for good classification performance. The statistical shape model (SSM) classifier achieved high classification metrics by re-parameterization of the 3D subject into its principal components, with 10 to 40 components achieving good performance, especially using naïve Bayes. Instead of selecting cephalometric parameters, this approach incorporated the full cranium into the classification which increased performance. The different morphing approaches showed little influence on the classification results. On the one hand, this implies classification robustness, but on the other hand, this might indicate that further improvements might be challenging, which falls in line with the observation that including the mirrored samples into the dataset did not improve performance. Classification using a convolutional neural network (CNN) scored the best performance metrics and showed attribution of the classification decision on areas of the head associated with craniosynostosis. Due to the 2D grid in which the distance maps are arranged, a low-resolution mapping could be designed which lead to a similar performance to the original images with a decrease in processing time. In the final experiment, clinical training data for the CNN could be replaced with data from multiple synthetic data sources without a performance drop. Overall, the CNN scored the best results, was the most flexible method in terms of data augmentation, and was able to classify clinical data without having seen a single clinical sample during training. As such, it is a likely contender for future diagnostic devices.

Two hypotheses had been proposed. The first hypothesis stated:

Hypothesis I

An F1-score of 0.95 or higher can be obtained by neural-network-based methods and non-neural-network-based methods.

During this thesis, multiple data-driven classification methods have been developed and compared. ResNet18 and GoogLeNet yielded F1-scores of 0.964 and 0.962 without data augmentation. The best non-neural-network-based approach was the SSM classification which yielded F1-score of 0.939, so the hypothesis has to be rejected. Including the mirrored samples for the SSM-based classifier could not improve classification performance. However, using neural networks was not a guarantee for a good classification performance, for example many CNNs and the feedforward neural network (FNN) did not yield an F1-score above 0.95 either. Especially the imbalanced dataset might have been a difficulty for the classifiers and their performance might therefore improve with a stratified dataset. However, the dataset distribution in this study is a snapshot from the real-world prevalences of craniosynostosis. As this was the largest classification study of craniosynostosis to date, it seems likely that the main findings hold up for other datasets as well.

The second hypothesis stated:

Hypothesis II

If trained on synthetic data, the F1-score of the classifier is at most 0.05 smaller compared to the classifier trained on clinical data.

The CNN classification approach was trained on multiple synthetic data sources and tested on clinical data. Compared with the training on clinical data, the F1-score slightly increased by 0.015 when using synthetic training data from an SSM and a generative adversarial network (GAN), so the hypothesis can be accepted. However, this is only valid for the two cases in which data was generated by an SSM and a GAN, and optionally using image-based principal component analysis (PCA). All other cases using synthetic data showed a performance drop of F1-score larger than 0.05. Their different approaches to modeling the underlying data distribution might be a reason why the combination of the two methods worked well.

Similar to the scientific hypothesis, four engineering milestones were initially proposed and achieved:

- **Systematically evaluate classification approaches:**
  The first systematic evaluation on the same dataset revealed that the proposed CNN-based classification on the 2D distance map outperformed competing approaches developed during this thesis and from the literature. This thesis also incorporated the first systematic evaluation of different classification

methods for craniosynostosis on the same dataset including methods from the literature.

- **Encode the 3D geometry into an anonymous representation:**
  The 3D-2D distance conversion proposed during this thesis proved to make head and face anonymous without losing classification ability. The 3D-2D conversion enabled additional benefits such as data augmentation during training and the usage of pre-trained CNNs. Multiple types of distance maps were developed, none of which was superior compared to the others. This suggests that *enabling* the usage of CNNs is actually more important for classification performance than the type of mathematical conversion. The FNN-based classification approach from the literature [13] also benefited from using the proposed 2D distance maps.
- **Synthesize pathology-specific data:**
  Using the pathology-specific submodels of an SSM, the first pathology-specific data synthesis for craniosynostosis was enabled. The SSM showed state-of-the-art performance. Combined with a GAN and an image-based PCA model, the SSM could be combined for a CNN-based classification based on synthetic data. In general, using multiple types of synthetic data sources lead to a better performance than relying on a single synthetic data source when using synthetic training data.
- **Increase data availability:**
  During this thesis, the first SSM of craniosynostosis patients was made publicly available including Python modules for data synthesis and 2D distance maps creation. As of September 2023, the model was downloaded more than 290 times. Additionally, the synthetic training data (GAN, SSM, and PCA) and the clinical 2D distance maps were made available which allows anyone to reproduce and improve the classification approach which adheres to the principles of the open science movement [172].

With the developed CNN-based classification, a versatile, accurate, and robust classification approach is now available which can exploit pre-trained CNNs and incorporate 2D data augmentation methods during training. In the long run, a diagnostic prediction tool for craniosynostosis requires trust of parents and physicians. To increase acceptance, the training of ML models could be performed on publicly available data. This data could be checked by clinical experts or even the public for quality and patient diversity in a collaborative process. As clinical data is unlikely to be made available in its raw format, the reliance on synthetic data is an important cornerstone for this approach. The synthetic classification study of this thesis and its data can be accessed and reproduced by anyone and is hopefully an incentive for other groups to publish synthetic or anonymized clinical data. As soon as other SSMs recorded from other hospitals are published, testing could be performed on other datasets to reduce potential biases in a single dataset. As the 3D-2D conversion

using 2D distance maps is easily applicable to other 3D meshes, a cross-domain 2D distance map dataset including CT-scans seems likely.

In this thesis, it was shown that classification of craniosynostosis can be performed on 3D surface scans. A CNN approach which outperformed competing approaches from the literature was proposed, which was able to classify clinical data even when it was only trained on synthetic data alone. These findings are important cornerstones for the development of data-driven and accurate radiation-free diagnostic tools for craniosynostosis. By translating the proposed methods into clinical practice, the treatment of head deformities can be organized in an open, effective, objective, and cost-effective manner.

Appendix A

Description of Alternative Morphing Algorithms

A.1 Iterative Coherent Point Drift

The iterative coherent point drift (ICPD) was proposed by [33] and iteratively uses nonrigid coherent point drift (CPD) [127] and consists of an initial morph using the Laplace-Beltrami regularized projection (LBRP) and a main loop in which the CPD alternates with a k-nearest-neighbors (kNN) selection. Some modifications were performed, e.g. instead of affine CPD, rigid CPD was used which increased robustness for the dataset. The algorithm is shown in Table A.1 with initial alignment of the reference shape $\mathbf{X}_\mathrm{r} \in \mathbb{R}^{p_\mathrm{X} \times 3}$ and target $\mathbf{Y} \in \mathbb{R}^{p_\mathrm{Y} \times 3}$. The main loop breaks out as soon as the change in each iteration of the nearest neighbors $\mathbf{id}_i$ of $\mathbf{X}_i$ is very small. Finally, a last LBRP step with low stiffness is performed.

Table A.1: Pseudocode of ICPD as used in this work, slighly adapted from the original authors [33].

1:	$\mathbf{X}_\mathrm{i} = \mathrm{LBRP}(\mathbf{X}_\mathrm{r}, \mathbf{Y})$ with high regularization
2:	$\mathbf{id}_\mathrm{i} = \mathrm{knnsearch}(\mathbf{Y}, \mathbf{X}_\mathrm{i})$
3:	Until $\mathsf{d}/\operatorname{len}(\mathbf{X}) < 0.01$:
4:	$\quad \mathbf{X}_{\mathrm{rig,i}} = \mathrm{cpdRigid}(\mathbf{X}_\mathrm{i}, \mathbf{Y}[\mathbf{id}_\mathrm{i}, :])$
5:	$\quad \mathbf{X}_\mathrm{i} = \mathrm{cpdNonrigid}(\mathbf{X}_{\mathrm{rig,i}}, \mathbf{Y}[\mathbf{id}_\mathrm{i}, :])$
6:	$\quad \mathbf{id}_\mathrm{old} = \mathbf{id}_\mathrm{i}$
7:	$\quad \mathbf{id}_\mathrm{i} = \mathrm{knnsearch}(\mathbf{Y}, \mathbf{X}_\mathrm{i})$
8:	$\quad \mathsf{d} = \mathrm{sum}(\mathrm{bool}(\mathrm{diff}(\mathbf{idx}_\mathrm{i}, \mathbf{idx}_\mathrm{old})))$
9:	$\mathbf{X} = \mathrm{LBRP}(\mathbf{X}_\mathrm{i}, \mathbf{Y})$ with low regularization

A.2 Nonrigid Optimal-Step Morphing Methods

The nonrigid iterative closest points affine (ICPA) and nonrigid iterative closest point translation (ICPT) methods were presented by [120] who base their work mostly on [130]. The core idea is to use an affine transformation for each point and locally regularize transformations of connected points. A stiffness term penalizes differences between transformations between adjacent nodes. A distance term controls how close the template vertices are transformed to the target points and a landmark term requires that the landmark points of template and target match each other. All three terms, stiffness term, distance term, and landmark term, are optimized simultaneously using an iterative approach starting with a high stiffness. For each stiffness, a correspondence search is performed and the optimal deformation with respect to the found correspondences is computed. As soon as the transformation changes little, the stiffness parameter is decreased and repeated for the reduced stiffness until convergence. For detailed explanations, the reader is referred to [120].

To be consistent with the notation in the original paper [120] for the description of the optimal-step nonrigid iterative closest points methods, the notation was changed. The n_p template points are expressed as $\mathbf{V} \in \mathbb{R}^{n_\mathrm{p} \times 3}$.

The unknown affine transformations are defined as $\mathbf{X} \in \mathbb{R}^{4n_\mathrm{p} \times 3}$. The full cost function can be expressed as

$$E(\mathbf{X}) = \alpha E_\mathrm{s}(\mathbf{X}) + E_\mathrm{d}(\mathbf{X}) + \beta E_\mathrm{l}(\mathbf{X}). \tag{A.1}$$

The stiffness term $E_\mathrm{s}(\mathbf{X})$ can be described as the Kronecker product $\otimes$ of the mesh topology matrix $\mathbf{M} \in \mathbb{R}^{n_\mathrm{e} \times n_\mathrm{p}}$ with n_e denoting the number of edges and n_p the number of points. The weight matrix $\mathbf{G} \in \mathbb{R}^{4\times4} = \mathrm{diag}(1, 1, 1, \gamma)$ between rotational and skew parts against translational parts [120] leads to:

$$E_\mathrm{s}(\mathbf{X}) = ||(\mathbf{M} \otimes \mathbf{G})\mathbf{X}||_F^2. \tag{A.2}$$

$\mathbf{M}$ describes the connections between neighboring vertices (the node-arc incidence matrix [173], in which for each edge r was set to $\mathbf{M}(r, i) = -1$ and $\mathbf{M}(r, j) = 1$). The distance term $E_\mathrm{d}(\mathbf{X})$ describes how close the displaced template vertices are to the target vertices and can be written as:

$$E_\mathrm{d}(\mathbf{X}) = ||\mathbf{W}(\mathbf{D}\mathbf{X} - \mathbf{U})||_F^2. \tag{A.3}$$

$\mathbf{W} \in \mathbb{R}^{n_\mathrm{p} \times n_\mathrm{p}}$ is a diagonal weighting matrix which allows assigning different weights to each transformation. The sparse displacement matrix $\mathbf{D} \in \mathbb{R}^{n_\mathrm{p} \times 4n_\mathrm{p}}$ is a diagonal matrix with the homogeneous points $v_i = [x_i, y_i, z_i, 1]^\mathsf{T}$ as its diagonal elements mapping the homogeneous template points to the respective affine transforms. $\mathbf{U} \in \mathbb{R}^{n_\mathrm{p} \times 3}$ denotes the found correspondences from the target points.

Finally, the landmark term $E_\mathrm{l}(\mathbf{X})$ is similar to the distance term while only the landmark points are considered:

$$E_\mathrm{l}(\mathbf{X}) = ||(\mathbf{D}_\mathrm{L}\mathbf{X} - \mathbf{U}_\mathrm{L})||_F^2. \tag{A.4}$$

The complete cost function for ICPA can be written as:

$$E(\mathbf{X}) = \left\| \begin{bmatrix} \alpha\mathbf{M} \otimes \mathbf{G} \\ \mathbf{W}\mathbf{D} \\ \beta\mathbf{D}_\mathrm{L} \end{bmatrix} \mathbf{X} - \begin{bmatrix} \mathbf{0} \\ \mathbf{W}\mathbf{U} \\ \mathbf{U}_\mathrm{L} \end{bmatrix} \right\|_F^2 \tag{A.5}$$

For the translation-only variant ICPT, the unknown transformations are defined as translations $\mathbf{X} \in \mathbb{R}^{n_\mathrm{p} \times 3}$. The cost function is changed accordingly:

$$E(\mathbf{X}) = \left\| \begin{bmatrix} \alpha\mathbf{M} \\ \mathbf{W}\mathbf{I}_{n_\mathrm{p}} \end{bmatrix} \mathbf{X} - \begin{bmatrix} \mathbf{0} \\ \mathbf{W}(\mathbf{U} - \mathbf{V}) \end{bmatrix} \right\|_F^2 \tag{A.6}$$

A.3 Hyperparameters for Template Morphing

Table A.2 lists the hyperparameters used in each method.

Table A.2: Hyperparameters used for the template morphing approaches.

Laplace-Beltrami regularized projection (LBRP) (notation of [33])	
Stiffness first morphing step	$\lambda_1 = 10$
Stiffness second morphing step	$\lambda_2 = 0.1$
Iterative coherent point drift (ICPD) (notation of [33])	
Stiffness first morph	$\lambda_1 = 10$
Iterative coherent point drift (ICPD)-Loop	For each iteration, perform first *cpdRigid*, then *cpdNonrigid*
cpdNonrigid smoothing weight:	3
cpdNonrigid tolerance	$1 \cdot 10^{-5}$
Exit condition	fewer than 1 % of nearest neighbors between iterations change
Stiffness second morph	$\lambda_2 = 0.1$ with Laplace matrix resulting from first morph
Nonrigid iterative closest points affine (ICPA) and nonrigid iterative closest point translation (ICPT) (notation of [120])	
Iterations	$\mathsf{n} = 80$
Stiffness parameter α in iteration n	$\alpha_\mathsf{n} = 10^8 \cdot 0.8^\mathsf{n}$
Landmark weight in iteration n β	if $\mathsf{n} \leq 50$ $\beta_\mathsf{n} = 1$, else $\beta_\mathsf{n} = 0$
Exit condition ϵ for each fixed stiffness α	$\epsilon < 100$
Valid normals for correspondence establishment φ	$\varphi < 45°$
Rotation weight γ	$\gamma = 1$

Appendix **B**

Additional Results of Cephalometric Multi-Height Classification

B.1 Quantitative Classification Results

The quantitative comparison of mean and standard deviations of all classifiers and feature extraction methods are displayed in Table B.1, which is a quantitative and more conventional display of the data compared to Fig. 7.7.

B.2 Cephalometric Multi-Height Classification Including Plagiocephaly

This section presents additional statistics of the cephalic index (CI) and cranial vault asymmetry index (CVAI) distribution and classification results if the plagiocephaly left and right groups are subdivided. Fig. B.1 shows an almost symmetric distribution of the values, while the classification results are shown in Fig. B.2 and Fig. B.3. While the trends observed in both experiments (with and without plagiocephaly) are similar, the classification performance even increased slightly with the plagiocephaly classes. It has to be noted however, that only the subjects in the dataset defined as left or right plagiocephaly were included in this study, so the number of samples is lower (484 instead of 496 with 12 fewer samples) and the results are therefore not quantitatively comparable.

Table B.1: Comparison of the classifiers and feature extraction methods. Displayed is cross validation mean ± standard deviation. This is the same as Fig. 7.7 in Chapter 7 but in a more conventional (but also less clear) manner.

Classifier	Features	Accuracy	G-mean	F1-score
SVM	Full Heights	0.927±0.026	0.321±0.395	0.769±0.091
SVM	Steps	0.929±0.024	0.377±0.377	0.775±0.076
SVM	Center Steps	0.927±0.027	0.394±0.402	0.776±0.103
SVM	SHAP	0.915±0.027	0.235±0.360	0.735±0.088
SVM	Middle	0.899±0.026	0.000±0.000	0.671±0.025
SVM	Max	0.696±0.024	0.000±0.000	0.369±0.021
LDA	Full Heights	0.613±0.045	0.299±0.301	0.510±0.061
LDA	Steps	**0.948±0.043**	**0.772±0.274**	**0.876±0.102**
LDA	Center Steps	0.940±0.042	0.683±0.350	0.846±0.109
LDA	SHAP	0.927±0.038	0.564±0.375	0.817±0.102
LDA	Middle	0.875±0.026	0.000±0.000	0.649±0.039
LDA	Max	0.734±0.036	0.000±0.000	0.441±0.079
NB	Full Heights	0.766±0.078	0.512±0.347	0.673±0.115
NB	Steps	0.780±0.071	0.361±0.363	0.644±0.111
NB	Center Steps	0.784±0.052	0.203±0.310	0.609±0.082
NB	SHAP	0.804±0.042	0.263±0.322	0.636±0.087
NB	Middle	0.796±0.022	0.000±0.000	0.550±0.032
NB	Max	0.655±0.023	0.000±0.000	0.376±0.070
KNN	Full Heights	0.901±0.027	0.378±0.380	0.745±0.097
KNN	Steps	0.893±0.031	0.296±0.364	0.719±0.088
KNN	Center Steps	0.928±0.037	0.558±0.370	0.804±0.103
KNN	SHAP	0.895±0.030	0.212±0.324	0.703±0.068
KNN	Middle	0.889±0.034	0.434±0.357	0.724±0.091
KNN	Max	0.683±0.027	0.042±0.125	0.436±0.059
DT	Full Heights	0.833±0.035	0.404±0.332	0.673±0.075
DT	Steps	0.837±0.060	0.421±0.346	0.670±0.116
DT	Center Steps	0.837±0.040	0.436±0.364	0.684±0.096
DT	SHAP	0.857±0.041	0.421±0.345	0.684±0.086
DT	Middle	0.822±0.040	0.070±0.211	0.599±0.075
DT	Max	0.708±0.039	0.101±0.202	0.455±0.079
RF	Full Heights	0.871±0.034	0.220±0.341	0.677±0.108
RF	Steps	0.855±0.025	0.204±0.316	0.648±0.083
RF	Center Steps	0.879±0.038	0.287±0.352	0.694±0.103
RF	SHAP	0.887±0.021	0.278±0.343	0.708±0.086
RF	Middle	0.855±0.039	0.068±0.204	0.633±0.065
RF	Max	0.708±0.028	0.000±0.000	0.425±0.060

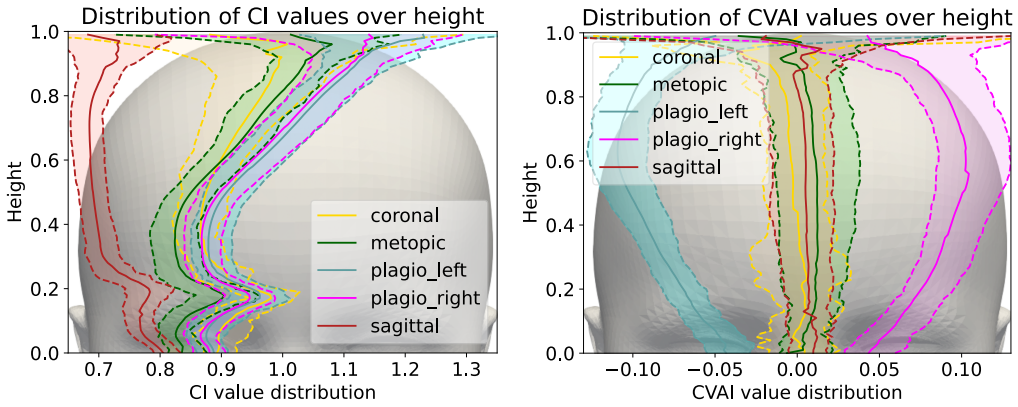


Figure B.1: Distribution of cephalic index (CI) and cranial vault asymmetry index (CVAI) including plagiocephaly with respect to the extraction height. The background visualizes the values on the extracted height on the mean shape of the control group, i.e., the extracted height is aligned with the background and the resulting distribution of CI and CVAI is placed on the x-axis. The mean value is shown as a solid line, the 25th and 75th percentiles are shown as dotted lines for each class.

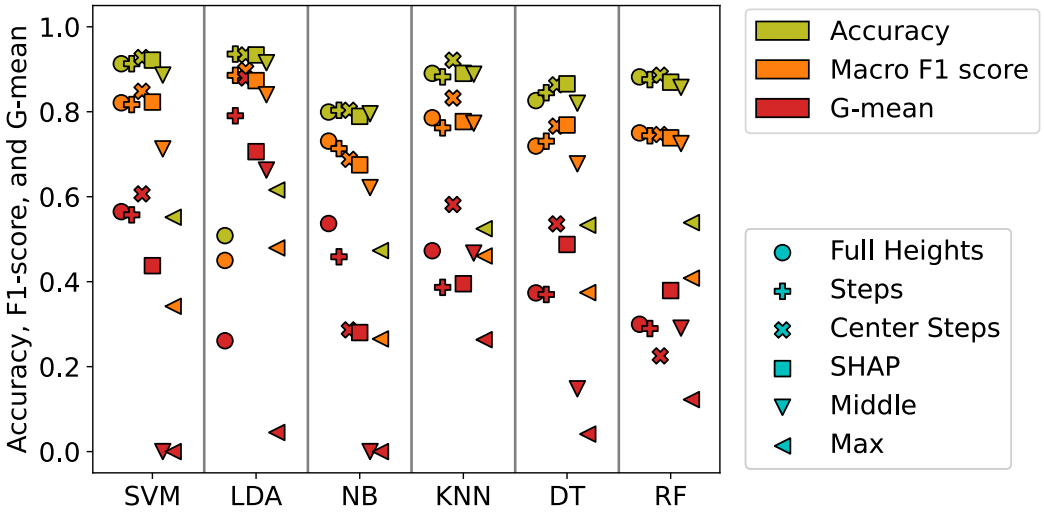


Figure B.2: Mean classification performance including plagiocephaly for different classification approaches and different classifiers using three-dimensional data for a 10-fold cross-validation approach. Each column shows a different classifier, colors show different mean performance metrics, while the symbols show different classification approaches.

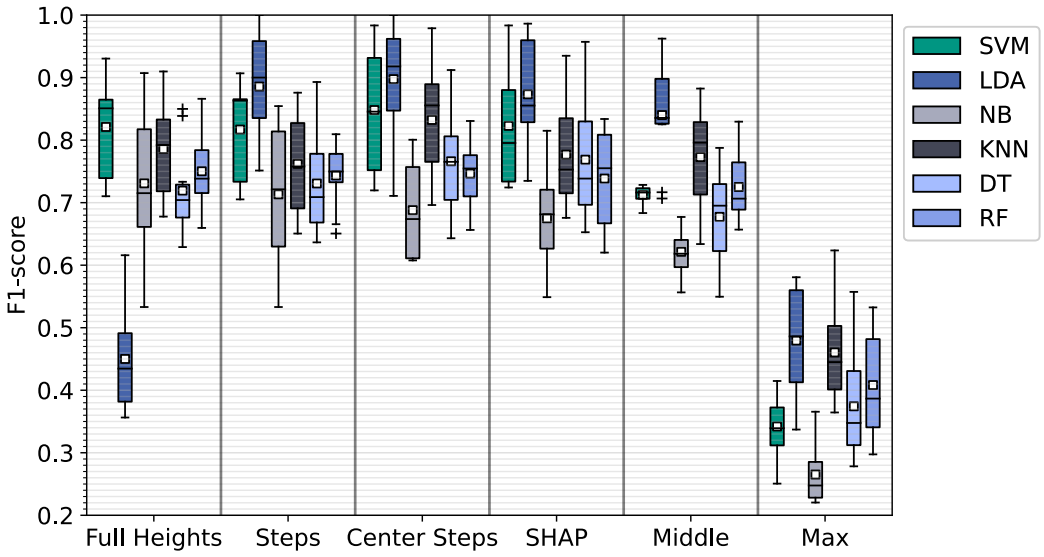


Figure B.3: Boxplots for the classification approaches including plagiocephaly using F1-score. Number of points on the head decreases monotonically from left to right.

Appendix C

Additional Results of Shape Model Classification

In this chapter, additional results for other morphing methods are presented. These are mainly classification plots of the other morphing methods, to undermine the claim in Chapter 8 that the classifiers and number of principal components show the same trend: Linear discriminant analysis (LDA), support vector machine (SVM), and naïve Bayes (NB) remain the most robust classifiers with regards to correctly classifying samples of the less represented classes in the dataset (indicated by the highest values of G-mean). The choice of principal components of 10 to 40 yields high performance metrics. The plots are represented in Figs. C.1, C.2, and C.3. Tab. C.1 shows the best performance metrics across all morphing methods.

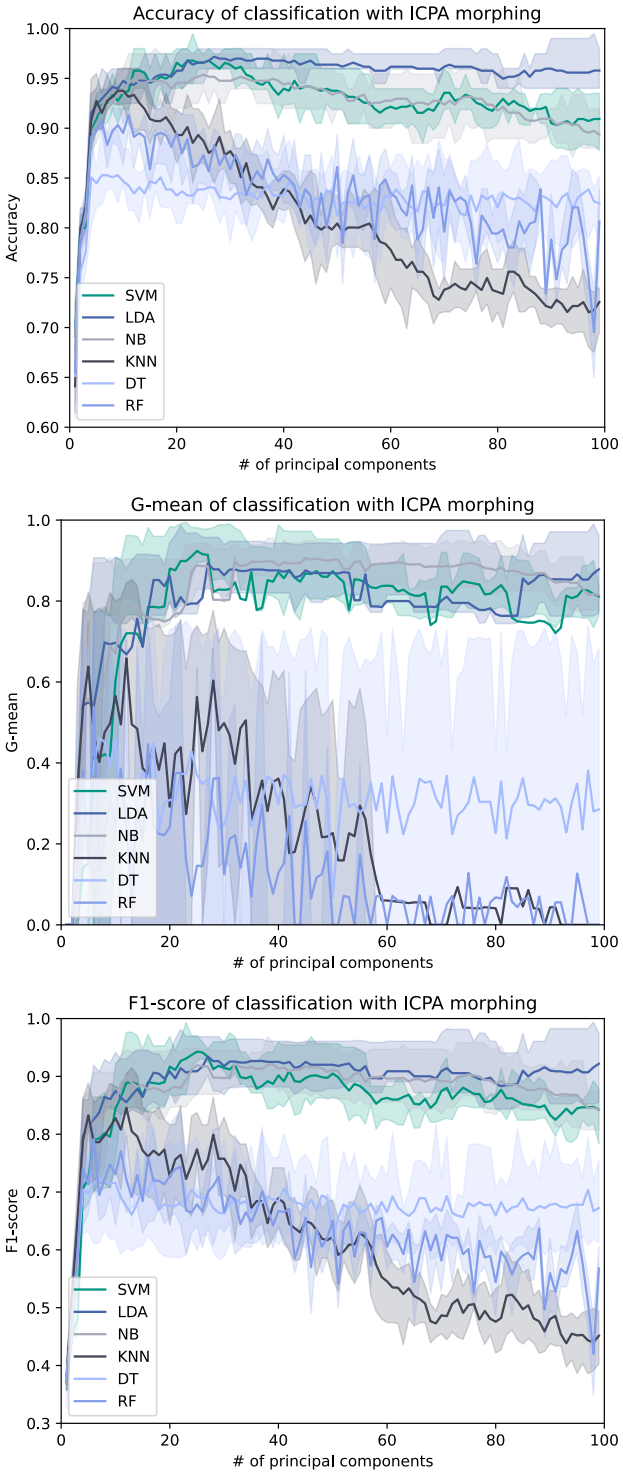


Figure C.1: Accuracy, G-mean, and F1-score as functions of the number of principal components used for the nonrigid iterative closest points affine (ICPA) classifier. Shown is the mean value and in lighter color the 25th and 75th percentiles.

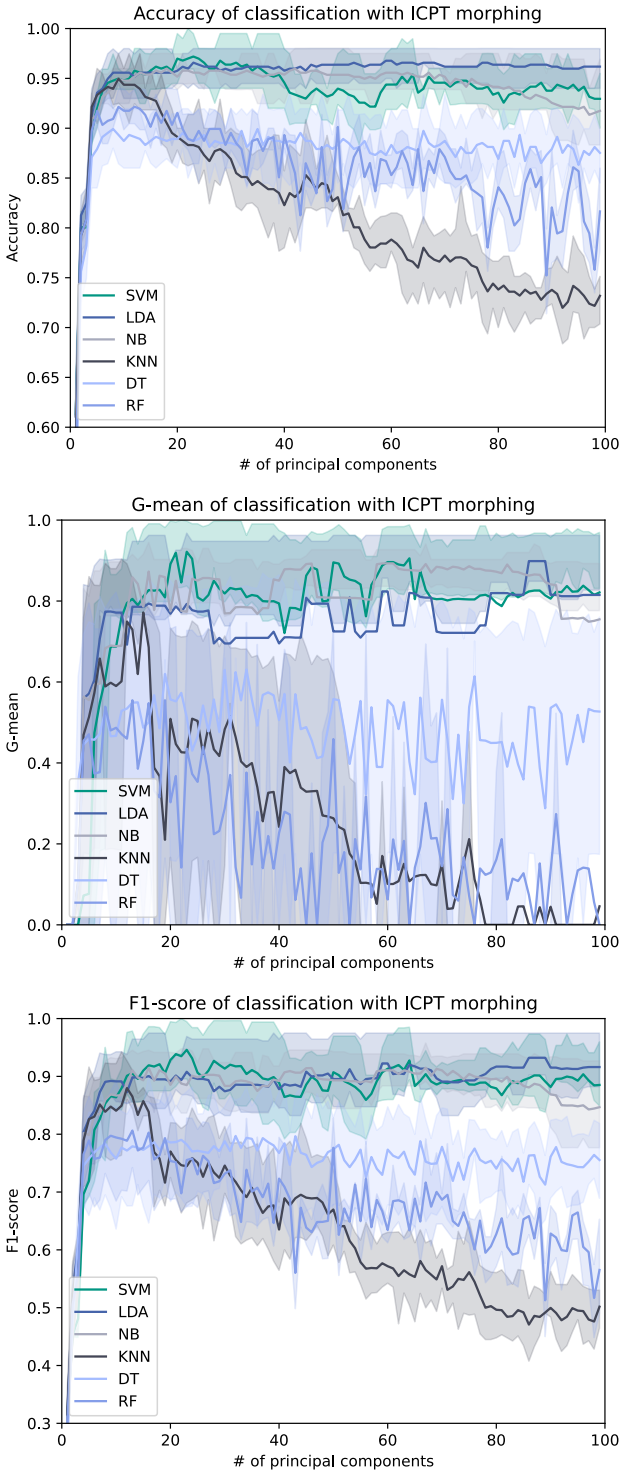


Figure C.2: Accuracy, G-mean, and F1-score as functions of the number of principal components used for the nonrigid iterative closest point translation (ICPT) classifier. Shown is the mean value and in lighter color the 25th and 75th percentiles.

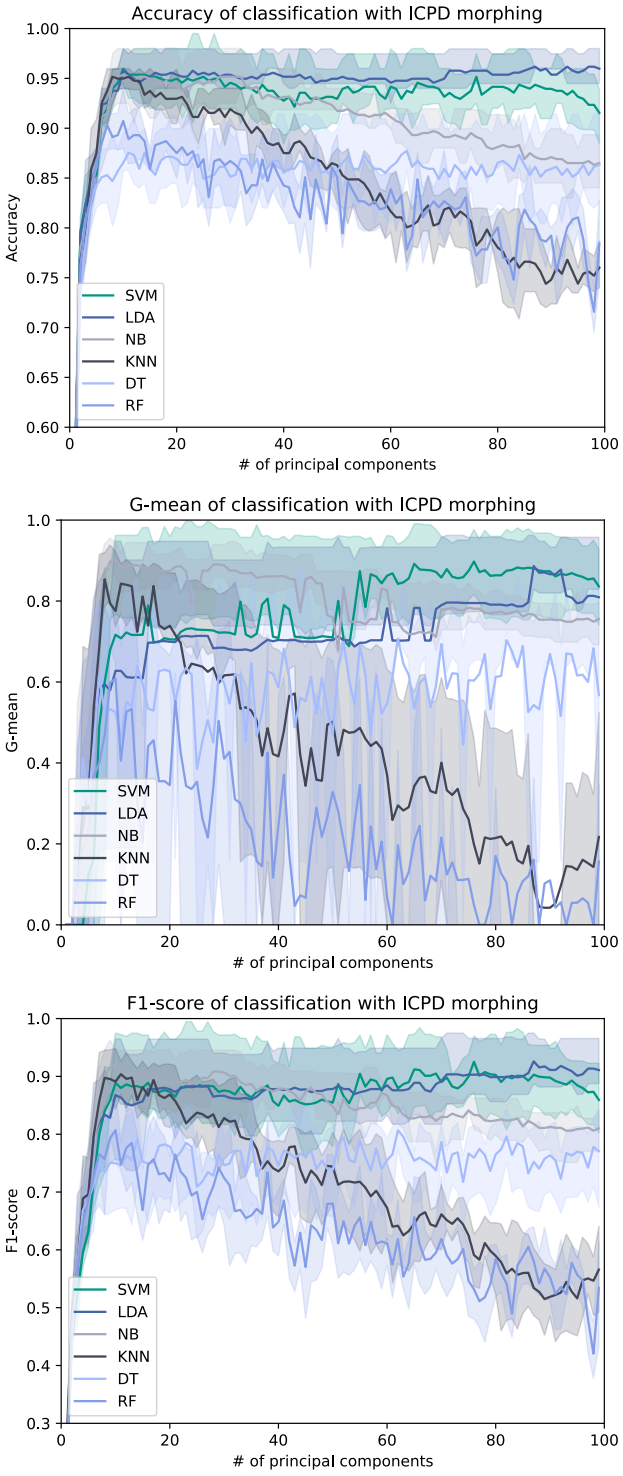


Figure C.3: Accuracy, G-mean, and F1-score as functions of the number of principal components used for the iterative coherent point drift (ICPD) classifier. Shown is the mean value and in lighter color the 25th and 75th percentiles.

Table C.1: Comparison of the classifiers on the cranium model across all morphing methods. Displayed is cross validation mean ± standard deviation.

Classifier	# components	Accuracy	G-mean	F1-score
Best individual classification run according to F1-score				
		LBRP		
SVM	(n=40)	0.958±0.029	0.918±0.086	0.931±0.058
LDA	(n=56)	0.962±0.032	0.718±0.365	0.891±0.101
NB	(n=12)	0.966±0.022	**0.926±0.053**	0.939±0.049
KNN	(n=7)	0.950±0.020	0.823±0.279	0.904±0.072
DT	(n=7)	0.851±0.047	0.667±0.235	0.773±0.076
RF	(n=9)	0.907±0.041	0.632±0.321	0.821±0.095
		ICPA		
SVM	(n=24)	0.966±0.016	0.924±0.074	0.943±0.036
LDA	(n=26)	**0.972±0.019**	0.893±0.078	0.934±0.047
NB	(n=31)	0.952±0.029	0.896±0.071	0.922±0.044
KNN	(n=11)	0.934±0.042	0.659±0.338	0.846±0.100
DT	(n=6)	0.853±0.054	0.461±0.381	0.724±0.119
RF	(n=5)	0.907±0.021	0.537±0.354	0.792±0.069
		ICPT		
SVM	(n=22)	0.972±0.022	0.922±0.106	**0.946±0.057**
LDA	(n=85)	0.964±0.028	0.899±0.086	0.932±0.054
NB	(n=66)	0.956±0.027	0.891±0.076	0.915±0.056
KNN	(n=11)	0.944±0.025	0.749±0.261	0.880±0.067
DT	(n=31)	0.897±0.034	0.624±0.321	0.797±0.088
RF	(n=12)	0.920±0.028	0.555±0.365	0.806±0.084
		ICPD		
SVM	(n=75)	0.952±0.030	0.898±0.099	0.925±0.061
LDA	(n=86)	0.962±0.030	0.886±0.086	0.926±0.056
NB	(n=9)	0.954±0.031	0.902±0.062	0.920±0.052
KNN	(n=10)	0.952±0.019	0.843±0.081	0.904±0.043
DT	(n=61)	0.873±0.065	0.707±0.254	0.807±0.086
RF	(n=9)	0.907±0.041	0.608±0.315	0.807±0.096

Appendix **D**

Description of Generative Adversarial Network Structure

This chapter contains the appendix of Chapter 10 and consists of the structure of the generative adversarial network (GAN) and information about the image artifacts which resulted from poor GAN designs.

D.1 GAN Artifacts

Fig. D.1 shows artifacts arising from only using transposed convolutional layers (`ConvTranspose2d`), using only up-scaling interpolation layers (`Interpolate`), or from large gradient penalties which prohibited training.

Figure D.1: Artifacts arising from a poor GAN design, displayed are four images, arranged in a 2×2 grid. From left to right: Deconvolution artifacts (checkerboard transposed convolution artifacts), interpolation (up-scaling) artifacts, and noise artifacts.

D.2 Network Structure

This is the GAN structure of generator and discriminator employed for the creation of the synthetic data. Output created using models' __str__ attribute.

```
Generator28(
  (embed): Embedding(4, 100)
  (gen): Sequential(
    (0): Sequential(
      (0): ConvTranspose2d(200, 256, kernel_size=(5, 5), stride=(1, 1),
                bias=False)
      (1): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=True)
      (2): ReLU(inplace=True)
    )
    (1): Sequential(
      (0): Interpolate(size=(8, 8),bilinear,align_corners=True)
      (1): BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=True)
      (2): ReLU(inplace=True)
    )
    (2): Sequential(
      (0): Conv2d(256, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1),
                bias=False)
      (1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=True)
      (2): ReLU(inplace=True)
    )
    (3): Sequential(
      (0): Interpolate(size=(15, 15),bilinear,align_corners=True)
      (1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=True)
      (2): ReLU(inplace=True)
    )
    (4): Sequential(
      (0): ConvTranspose2d(128, 128, kernel_size=(3, 3), stride=(1, 1),
                bias=False)
      (1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=True)
      (2): ReLU(inplace=True)
    )
    (5): Sequential(
      (0): Interpolate(size=(30, 30),bilinear,align_corners=True)
      (1): BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=True)
      (2): ReLU(inplace=True)
    )
    (6): Conv2d(128, 1, kernel_size=(3, 3), stride=(1, 1),
                bias=False)
    (7): Tanh()
```

```
  )
)

Discriminator28(
  (net): Sequential(
    (0): Sequential(
      (0): Conv2d(2, 32, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1),
                bias=False)
      (1): InstanceNorm2d(32, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=False)
      (2): LeakyReLU(negative_slope=0.2)
    )
    (1): Sequential(
      (0): Conv2d(32, 128, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1),
                bias=False)
      (1): InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=False)
      (2): LeakyReLU(negative_slope=0.2)
    )
    (2): Sequential(
      (0): Conv2d(128, 256, kernel_size=(5, 5), stride=(2, 2), padding=(1, 1),
                bias=False)
      (1): InstanceNorm2d(256, eps=1e-05, momentum=0.1, affine=True,
                track_running_stats=False)
      (2): LeakyReLU(negative_slope=0.2)
    )
    (3): Conv2d(256, 1, kernel_size=(3, 3), stride=(1, 1))
  )
  (embed): Embedding(4, 784)
)
```

References

[1] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019. doi:10.1038/s41591-018-0300-7

[2] C. Sabet, A. Hammond, N. Ravid, M. S. Tong, and F. C. Stanford, "Harnessing big data for health equity through a comprehensive public database and data collection framework," *npj Digital Medicine*, vol. 6, pp. 1–2, 2023. doi:10.1038/s41746-023-00844-5

[3] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nature Biomedical Engineering*, vol. 6, pp. 1330–1345, 2022. doi:10.1038/s41551-022-00898-y

[4] J. Banerjee, J. N. Taroni, R. J. Allaway, D. V. Prasad, J. Guinney, and C. Greene, "Machine learning in rare disease," *Nature Methods*, vol. 20, pp. 803–814, 2023. doi:10.1038/s41592-023-01886-z

[5] D. Renier, C. Sainte-Rose, D. Marchac, and J.-F. Hirsch, "Intracranial pressure in craniostenosis," *Journal of Neurosurgery*, vol. 57, pp. 370–377, 1982. doi:10.3171/jns.1982.57.3.0370

[6] K. A. Kapp-Simon, M. L. Speltz, M. L. Cunningham, P. K. Patel, and T. Tomita, "Neurodevelopment of children with single suture craniosynostosis: A review," *Child's Nervous System*, vol. 23, pp. 269–281, 2007. doi:10.1007/s00381-006-0251-z

[7] D. B. Becker, J. D. Petersen, A. A. Kane, M. M. Cradock, T. K. Pilgram, and J. L. Marsh, "Speech, Cognitive, and Behavioral Outcomes in Nonsyndromic Craniosynostosis," *Plastic and Reconstructive Surgery*, vol. 116, pp. 400–407, 2005. doi:10.1097/01.prs.0000172763.71043.b8

[8] M. Engel, G. Castrillon-Oberndorfer, J. Hoffmann, and C. Freudlsperger, "Value of preoperative imaging in the diagnostics of isolated metopic suture synostosis: A risk–benefit analysis," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 65, pp. 1246–1251, 2012. doi:10.1016/j.bjps.2012.03.038

[9] J. Panchal and V. Uttchin, "Management of Craniosynostosis," *Plastic and Reconstructive Surgery*, vol. 111, pp. 2032–2048, 2003. doi:10.1097/01.PRS.0000056839.94034.47

[10] G. Cacciaguerra, M. Palermo, L. Marino, et al., "The Evolution of the Role of Imaging in the Diagnosis of Craniosynostosis: A Narrative Review," *Children*, vol. 8, p. 727, 2021. doi:10.3390/children8090727

[11] C. Mertens, E. Wessel, M. Berger, et al., "The value of three-dimensional photogrammetry in isolated sagittal synostosis: Impact of age and surgical technique on intracranial volume and cephalic index-a retrospective cohort study," *Journal of Cranio-Maxillofacial Surgery*, vol. 45, pp. 2010–2016, 2017. doi:10.1016/j.jcms.2017.09.019

[12] S. L. Boulet, S. A. Rasmussen, and M. A. Honein, "A population-based study of craniosynostosis in metropolitan Atlanta, 1989–2003," *American Journal of Medical Genetics Part A*, vol. 146A, pp. 984–991, 2008. doi:10.1002/ajmg.a.32208

[13] G. de Jong, E. Bijlsma, J. Meulstee, et al., "Combining deep learning with 3D stereophotogrammetry for craniosynostosis diagnosis," *Scientific Reports*, vol. 10, p. 15346, 2020. doi:10.1038/s41598-020-72143-y

[14] N. F. Banner, "The human side of health data," *Nature Medicine*, vol. 26, pp. 995–995, 2020. doi:10.1038/s41591-020-0838-z

[15] T. Elarjani, O. T. Almutairi, M. Alhussinan, et al., "Bibliometric analysis of the top 100 most cited articles on craniosynostosis," *Child's Nervous System*, vol. 37, pp. 587–597, 2021. doi:10.1007/s00381-020-04858-2

[16] R. Iping, A. M. Cohen, T. Abdel Alim, et al., "A bibliometric overview of craniosynostosis research development," *European Journal of Medical Genetics*, vol. 64, p. 104224, 2021. doi:10.1016/j.ejmg.2021.104224

[17] M. W. Vannier, C. F. Hildebolt, J. L. Marsh, et al., "Craniosynostosis: Diagnostic value of three-dimensional CT reconstruction." *Radiology*, vol. 173, pp. 669–673, 1989. doi:10.1148/radiology.173.3.2813770

[18] L. Massimi, F. Bianchi, P. Frassanito, R. Calandrelli, G. Tamburrini, and M. Caldarelli, "Imaging in craniosynostosis: When and what?" *Child's Nervous System*, vol. 35, pp. 2055–2069, 2019. doi:10.1007/s00381-019-04278-x

[19] J. Weinzweig, R. E. Kirschner, A. Farley, et al., "Metopic Synostosis: Defining the Temporal Sequence of Normal Suture Fusion and Differentiating It from Synostosis on the Basis of Computed Tomography Images," *Plastic and Reconstructive Surgery*, vol. 112, pp. 1211–1218, 2003. doi:10.1097/01.PRS.0000080729.28749.A3

[20] J. R. Marcus, L. F. Domeshek, A. M. Loyd, et al., "Use of a Three-Dimensional, Normative Database of Pediatric Craniofacial Morphology for Modern Anthropometric Analysis," *Plastic and Reconstructive Surgery*, vol. 124, pp. 2076–2084, 2009. doi:10.1097/PRS.0b013e3181bf7e1b

[21] A. Fabijańska and T. Węgliński, "The quantitative assessment of the pre- and postoperative craniosynostosis using the methods of image analysis," *Computerized Medical Imaging and Graphics*, vol. 46, pp. 153–168, 2015. doi:10.1016/j.compmedimag.2015.05.005

[22] C. S. Mendoza, N. Safdar, E. Myers, T. Kittisarapong, G. F. Rogers, and M. G. Linguraru, "Computer-Based Quantitative Assessment of Skull Morphology for Craniosynostosis," in *Clinical Image-Based Procedures. From Planning to Intervention*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, vol. 7761, pp. 98–105. doi:10.1007/978-3-642-38079-2_13

[23] R. Calandrelli, F. Pilato, L. Massimi, M. Panfili, C. Di Rocco, and C. Colosimo, "The unseen third dimension: A novel approach for assessing head shape severity in infants with isolated sagittal synostosis," *Child's Nervous System*, vol. 35, pp. 1351–1356, 2019. doi:10.1007/s00381-019-04246-5

[24] O. D. M. Kronig, S. A. J. Kronig, H. A. Vrooman, et al., "Introducing a new method for classifying skull shape abnormalities related to craniosynostosis," *European Journal of Pediatrics*, vol. 179, pp. 1569–1577, 2020. doi:10.1007/s00431-020-03643-2

[25] S. Ruiz-Correa, R. Sze, H. Lin, L. Shapiro, M. Speltz, and M. Cunningham, "Classifying Craniosynostosis Deformations by Skull Shape Imaging," in *18th IEEE Symposium on*

Computer-Based Medical Systems (CBMS'05). Dublin, Ireland: IEEE, 2005, pp. 335–340. doi:10.1109/CBMS.2005.42

[26] S. Ruiz-Correa, D. Gatica-Perez, H. J. Lin, L. G. Shapiro, and R. Sze, "A Bayesian hierarchical model for classifying craniofacial malformations from CT imaging," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vancouver, BC: IEEE, 2008, pp. 4063–4069. doi:10.1109/IEMBS.2008.4650102

[27] L. You, G. Zhang, W. Zhao, M. G. R, L. David, and X. Zhou, "Automated Sagittal Craniosynostosis Classification from CT Images Using Transfer Learning," *Clinics in Surgery*, vol. 5, p. 2746, 2020.

[28] L. You, Y. Deng, G. Zhang, et al., "A novel sagittal craniosynostosis classification system based on multi-view learning algorithm," *Neural Computing and Applications*, vol. 34, pp. 14 427–14 434, 2022. doi:10.1007/s00521-022-07310-5

[29] C. S. Mendoza, N. Safdar, K. Okada, E. Myers, G. F. Rogers, and M. G. Linguraru, "Personalized assessment of craniosynostosis via statistical shape modeling," *Medical Image Analysis*, vol. 18, pp. 635–646, 2014. doi:10.1016/j.media.2014.02.008

[30] J. Meulstee, L. Verhamme, W. Borstlap, et al., "A new method for three-dimensional evaluation of the cranial shape and the automatic identification of craniosynostosis using 3D stereophotogrammetry," *International Journal of Oral and Maxillofacial Surgery*, vol. 46, pp. 819–826, 2017. doi:10.1016/j.ijom.2017.03.017

[31] P. Heutinck, P. Knoops, N. R. Florez, et al., "Statistical shape modelling for the analysis of head shape variations," *Journal of Cranio-Maxillofacial Surgery*, vol. 49, pp. 449–455, 2021. doi:10.1016/j.jcms.2021.02.020

[32] H. Dai, N. Pears, and C. Duncan, "A 2D Morphable Model of Craniofacial Profile and Its Application to Craniosynostosis," in *Medical Image Understanding and Analysis*. Cham: Springer International Publishing, 2017, vol. 723, pp. 731–742. doi:10.1007/978-3-319-60964-5_64

[33] H. Dai, N. Pears, W. Smith, and C. Duncan, "Statistical Modeling of Craniofacial Shape and Texture," *International Journal of Computer Vision*, vol. 128, pp. 547–571, 2020. doi:10.1007/s11263-019-01260-7

[34] S. Lanche, T. A. Darvann, H. Ólafsdóttir, et al., "A Statistical Model of Head Asymmetry in Infants with Deformational Plagiocephaly," in *Image Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, vol. 4522, pp. 898–907. doi:10.1007/978-3-540-73040-8_91

[35] G. de Jong, T. Maal, and H. Delye, "The computed cranial focal point," *Journal of Cranio-Maxillofacial Surgery*, vol. 43, pp. 1737–1742, 2015. doi:10.1016/j.jcms.2015.08.023

[36] G. De Jong, M. Tolhuisen, J. Meulstee, et al., "Radiation-free 3D head shape and volume evaluation after endoscopically assisted strip craniectomy followed by helmet therapy for trigonocephaly," *Journal of Cranio-Maxillofacial Surgery*, vol. 45, pp. 661–671, 2017. doi:10.1016/j.jcms.2017.02.007

[37] C. A. Beaumont, P. G. Knoops, A. Borghi, et al., "Three-dimensional surface scanners compared with standard anthropometric measurements for head shape," *Journal of Cranio-Maxillofacial Surgery*, vol. 45, pp. 921–927, 2017. doi:10.1016/j.jcms.2017.03.003

[38] S. Agarwal, R. R. Hallac, R. Mishra, C. Li, O. Daescu, and A. Kane, "Image Based Detection of Craniofacial Abnormalities using Feature Extraction by Classical Convolutional Neural Network," in *2018 IEEE 8th International Conference on Computational*

Advances in Bio and Medical Sciences (ICCABS). Las Vegas, NV: IEEE, 2018, pp. 1–6. doi:10.1109/ICCABS.2018.8541948

[39] S. Agarwal, R. R. Hallac, O. Daescu, and A. Kane, "Classification of Craniosynostosis Images by Vigilant Feature Extraction," in *Advances in Computer Vision and Computational Biology*. Cham: Springer International Publishing, 2021, pp. 293–306. doi:10.1007/978-3-030-71051-4_23

[40] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, pp. 452–454, 2016. doi:10.1038/533452a

[41] the Precise4Q consortium, J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, p. 310, 2020. doi:10.1186/s12911-020-01332-6

[42] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019. doi:10.1038/s42256-019-0048-x

[43] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"," *AI Magazine*, vol. 38, pp. 50–57, 2017. doi:10.1609/aimag.v38i3.2741

[44] T. W. Sadler, *Langman's medical embryology*, 11th ed. Philadelphia: Wolters Kluwer, 2010.

[45] J. Sobotta, *Atlas and textbook of human anatomy*, 1905.

[46] S. Standring, *Gray's Anatomy: The Anatomical Basis of Clinical Practice*. Elsevier Limited, 2016.

[47] S. Harth, M. Obert, F. Ramsthaler, C. Reuß, H. Traupe, and M. A. Verhoff, "Estimating age by assessing the ossification degree of cranial sutures with the aid of Flat-Panel-CT," *Legal Medicine*, vol. 11, pp. S186–S189, 2009. doi:10.1016/j.legalmed.2009.01.091

[48] L. French, I. T. Jackson, and L. Melton, "A population-based study of craniosynostosis," *Journal of Clinical Epidemiology*, vol. 43, pp. 69–73, 1990. doi:10.1016/0895-4356(90)90058-W

[49] A. Shuper, "The Incidence of Isolated Craniosynostosis in the Newborn Infant," *Archives of Pediatrics & Adolescent Medicine*, vol. 139, p. 85, 1985. doi:10.1001/archpedi.1985.02140030091038

[50] E. Tønne, B. J. Due-Tønnessen, U. Wiig, et al., "Epidemiology of craniosynostosis in Norway," *Journal of Neurosurgery: Pediatrics*, vol. 26, pp. 68–75, 2020. doi:10.3171/2020.1.PEDS2051

[51] J. S. Lee and J. W. Yu, "Craniosynostosis Surgery," in *The History of Maxillofacial Surgery*. Cham: Springer International Publishing, 2022, pp. 367–390. doi:10.1007/978-3-030-89563-1_20

[52] R. Virchow, "Über den Cretinismus, namentlich in Franken, und über pathologische Schädelformen," *Verh Phys Med Ges Würz*, vol. 2, pp. 230–270, 1851.

[53] J. A. Persing, J. A. Jane, and M. Shaffrey, "Virchow and the Pathogenesis of Craniosynostosis: A Translation of His Original Work," *Plastic and Reconstructive Surgery*, vol. 83, pp. 738–742, 1989. doi:10.1097/00006534-198904000-00025

[54] B. J. Slater, K. A. Lenton, M. D. Kwan, D. M. Gupta, D. C. Wan, and M. T. Longaker, "Cranial Sutures: A Brief Review," *Plastic and Reconstructive Surgery*, vol. 121, pp. 170e–178e, 2008. doi:10.1097/01.prs.0000304441.99483.97

[55] A. O. Wilkie, D. Johnson, and S. A. Wall, "Clinical genetics of craniosynostosis," *Current Opinion in Pediatrics*, vol. 29, pp. 622–628, 2017. doi:10.1097/MOP.0000000000000542

[56] A. K. Coussens, C. R. Wilkinson, I. P. Hughes, et al., "Unravelling the molecular control of calvarial suture fusion in children with craniosynostosis," *BMC Genomics*, vol. 8, p. 458, 2007. doi:10.1186/1471-2164-8-458

[57] B. W. Alderman, C. M. Bradley, C. Greene, S. K. Fernbach, and A. E. Barón, "Increased risk of craniosynostosis with maternal cigarette smoking during pregnancy," *Teratology*, vol. 50, pp. 13–18, 1994. doi:10.1002/tera.1420500103

[58] M. S. Dias, T. Samson, E. B. Rizk, L. S. Governale, J. T. Richtsmeier, and section on neurologic surgery, section on plastic and reconstructive surgery, "Identifying the Misshapen Head: Craniosynostosis and Related Disorders," *Pediatrics*, vol. 146, p. e2020015511, 2020. doi:10.1542/peds.2020-015511

[59] D. T. Gault, D. Renier, D. Marchac, and B. M. Jones, "Intracranial Pressure and Intracranial Volume in Children with Craniosynostosis," *Plastic and Reconstructive Surgery*, vol. 90, pp. 377–381, 1992. doi:10.1097/00006534-199209000-00003

[60] K. A. Eley, S. R. Watt-Smith, F. Sheerin, and S. J. Golding, ""Black Bone" MRI: A potential alternative to CT with three-dimensional reconstruction of the craniofacial skeleton in the diagnosis of craniosynostosis," *European Radiology*, vol. 24, pp. 2417–2426, 2014. doi:10.1007/s00330-014-3286-7

[61] J. Regelsberger, G. Delling, K. Helmke, et al., "Ultrasound in the Diagnosis of Craniosynostosis," *Journal of Craniofacial Surgery*, vol. 17, pp. 623–625, 2006. doi:10.1097/00001665-200607000-00002

[62] A. Saarikko, E. Mellanen, L. Kuusela, et al., "Comparison of Black Bone MRI and 3D-CT in the preoperative evaluation of patients with craniosynostosis," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, vol. 73, pp. 723–731, 2020. doi:10.1016/j.bjps.2019.11.006

[63] B. F. Judy, J. W. Swanson, W. Yang, et al., "Intraoperative intracranial pressure monitoring in the pediatric craniosynostosis population," *Journal of Neurosurgery: Pediatrics*, vol. 22, pp. 475–480, 2018. doi:10.3171/2018.5.PEDS1876

[64] N. Bannink, E. Nout, E. Wolvius, H. Hoeve, K. Joosten, and I. Mathijssen, "Obstructive sleep apnea in children with syndromic craniosynostosis: Long-term respiratory outcome of midface advancement," *International Journal of Oral and Maxillofacial Surgery*, vol. 39, pp. 115–121, 2010. doi:10.1016/j.ijom.2009.11.021

[65] J. A. Fearon, R. A. Ruotolo, and J. C. Kolar, "Single Sutural Craniosynostoses: Surgical Outcomes and Long-Term Growth," *Plastic and Reconstructive Surgery*, vol. 123, pp. 635–642, 2009. doi:10.1097/PRS.0b013e318195661a

[66] S. Nagaraja, P. Anslow, and B. Winter, "Craniosynostosis," *Clinical Radiology*, vol. 68, pp. 284–292, 2013. doi:10.1016/j.crad.2012.07.005

[67] S. A. G. Roberts, J. D. Symonds, R. Chawla, E. Toman, J. Bishop, and G. A. Solanki, "Positional plagiocephaly following ventriculoperitoneal shunting in neonates and infancy—how serious is it?" *Child's Nervous System*, vol. 33, pp. 275–280, 2017. doi:10.1007/s00381-016-3275-z

[68] J. Persing, H. James, J. Swanson, et al., "Prevention and Management of Positional Skull Deformities in Infants," *Pediatrics*, vol. 112, pp. 199–202, 2003. doi:10.1542/peds.112.1.199

[69] J. Laughlin, T. G. Luerssen, M. S. Dias, and the Committee on Practice and Ambulatory Medicine, Section on Neurological Surgery, "Prevention and Management of Positional Skull Deformities in Infants," *Pediatrics*, vol. 128, pp. 1236–1241, 2011. doi:10.1542/peds.2011-2220

[70] C. Freudlsperger, S. Steinmacher, D. Saure, et al., "Impact of severity and therapy onset on helmet therapy in positional plagiocephaly," *Journal of Cranio-Maxillofacial Surgery*, vol. 44, pp. 110–115, 2016. doi:10.1016/j.jcms.2015.11.016

[71] C. A. Callejas Pastor, I.-Y. Jung, S. Seo, S. B. Kwon, Y. Ku, and J. Choi, "Two-Dimensional Image-Based Screening Tool for Infants with Positional Cranial Deformities: A Machine Learning Approach," *Diagnostics*, vol. 10, p. 495, 2020. doi:10.3390/diagnostics10070495

[72] T. Koizumi, Y. Komuro, K. Hashizume, and A. Yanai, "Cephalic Index of Japanese Children With Normal Brain Development," *Journal of Craniofacial Surgery*, vol. 21, pp. 1434–1437, 2010. doi:10.1097/SCS.0b013e3181ecc2f3

[73] A. A. Waitzman, J. C. Posnick, D. C. Armstrong, and G. E. Pron, "Craniofacial Skeletal Measurements Based on Computed Tomography: Part II. Normal Values and Growth Trends," *The Cleft Palate-Craniofacial Journal*, vol. 29, pp. 118–128, 1992. doi:10.1597/1545-1569_1992_029_0118_csmboc_2.3.co_2

[74] A. A. Waitzman, J. C. Posnick, D. C. Armstrong, and G. E. Pron, "Craniofacial Skeletal Measurements Based on Computed Tomography: Part I. Accuracy and Reproducibility," *The Cleft Palate-Craniofacial Journal*, vol. 29, pp. 112–117, 1992. doi:10.1597/1545-1569_1992_029_0112_csmboc_2.3.co_2

[75] N. A. Pickersgill, G. B. Skolnick, S. D. Naidoo, M. D. Smyth, and K. B. Patel, "Regression of cephalic index following endoscopic repair of sagittal synostosis," *Journal of Neurosurgery: Pediatrics*, vol. 23, pp. 54–60, 2019. doi:10.3171/2018.7.PEDS18195

[76] H. Nam, N. Han, M. J. Eom, M. Kook, and J. Kim, "Cephalic Index of Korean Children With Normal Brain Development During the First 7 Years of Life Based on Computed Tomography," *Annals of Rehabilitation Medicine*, vol. 45, pp. 141–149, 2021. doi:10.5535/arm.20235

[77] W. Likus, G. Bajor, K. Gruszczyńska, et al., "Cephalic Index in the First Three Years of Life: Study of Children with Normal Brain Development Based on Computed Tomography," *The Scientific World Journal*, vol. 2014, pp. 1–6, 2014. doi:10.1155/2014/502836

[78] M. A. Musa, A. D. Zagga, M. Danfulani, A. A. Tadros, and A. Hamid, "Cranial index of children with normal and abnormal brain development in Sokoto, Nigeria: A comparative study," *Journal of Neurosciences in Rural Practice*, vol. 5, pp. 139–143, 2014. doi:10.4103/0976-3147.131655

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

[80] Q. Li, H. Peng, J. Li, et al., "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, pp. 1–41, 2022. doi:10.1145/3495162

[81] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, pp. 210–229, 1959. doi:10.1147/rd.33.0210

[82] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees." *Biometrics*, vol. 40, p. 874, 1984. doi:10.2307/2530946

[83] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. doi:10.1023/A:1010933404324

[84] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967. doi:10.1109/TIT.1967.1053964

[85] H. Zhang, "The optimality of naive bayes," in *The Florida AI Research Society*, 2004.

[86] O. Kupervasser, "The mysterious optimality of Naive Bayes: Estimation of the probability in the system of "classifiers"," *Pattern Recognition and Image Analysis*, vol. 24, pp. 1–10, 2014. doi:10.1134/S1054661814010088

[87] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936. doi:10.1111/j.1469-1809.1936.tb02137.x

[88] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. doi:10.1007/BF00994018

[89] A. Paszke, S. Gross, F. Massa, et al., "Pytorch: an imperative style, high-performance deep learning library," in *Advances in neural information processing systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[90] O. Russakovsky, J. Deng, H. Su, et al., "ImageNet Large Scale Visual Recognition Challenge," 2014. doi:10.48550/ARXIV.1409.0575

[91] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, and C. Ou, "Automatic ECG Classification Using Continuous Wavelet Transform and Convolutional Neural Network," *Entropy*, vol. 23, p. 119, 2021. doi:10.3390/e23010119

[92] C. Song, Y. Huang, W. Wang, and L. Wang, "CASIA-E: A Large Comprehensive Dataset for Gait Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2022. doi:10.1109/TPAMI.2022.3183288

[93] N. Pilia, S. Schuler, M. Rees, et al., "Non-invasive Localization of the Ventricular Excitation Origin Without Patient-specific Geometries Using Deep Learning," 2022. doi:10.48550/arXiv.2209.08095

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, 2017. doi:10.1145/3065386

[95] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010. doi:10.1109/TKDE.2009.191

[96] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015. doi:10.1038/nature14539

[97] J. Bennett, S. Lanning, et al., "The netflix prize," in *Proceedings of KDD Cup and Workshop*, vol. 2007. New York, 2007, p. 35.

[98] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proceedings of the 14th International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[99] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," 2014. doi:10.48550/arXiv.1406.2661

[100] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," 2016. doi:10.48550/arXiv.1511.06434

[101] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," 2019. doi:10.48550/arXiv.1812.04948

[102] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep Photo Style Transfer," 2017. doi:10.48550/arXiv.1703.07511

[103] J.-Y. Zhu, R. Zhang, D. Pathak, et al., "Toward Multimodal Image-to-Image Translation," 2018. doi:10.48550/arXiv.1711.11586

[104] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," 2016. doi:10.48550/arXiv.1604.07379

[105] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," 2014. doi:10.48550/arXiv.1411.1784

[106] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017. doi:10.48550/arXiv.1701.07875

[107] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," 2017. doi:10.48550/arXiv.1704.00028

[108] T. Albrecht, M. Lüthi, T. Gerig, and T. Vetter, "Posterior shape models," *Medical Image Analysis*, vol. 17, pp. 959–973, 2013. doi:10.1016/j.media.2013.05.010

[109] M. Luthi, T. Gerig, C. Jud, and T. Vetter, "Gaussian Process Morphable Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 1860–1873, 2018. doi:10.1109/TPAMI.2017.2739743

[110] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*. Not Known: ACM Press, 1999, pp. 187–194. doi:10.1145/311535.311556

[111] D. G. Kendall, "Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces," *Bulletin of the London Mathematical Society*, vol. 16, pp. 81–121, 1984. doi:10.1112/blms/16.2.81

[112] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Training Models of Shape from Sets of Examples," in *BMVC92*. London: Springer London, 1992, pp. 9–18. doi:10.1007/978-1-4471-3201-1_2

[113] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models-Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995. doi:10.1006/cviu.1995.1004

[114] T. Gerig, A. Morel-Forster, C. Blumer, et al., "Morphable Face Models - An Open Framework," *arXiv:1709.08398 [cs]*, 2017.

[115] H. Dai, N. Pears, W. Smith, and C. Duncan, "A 3D Morphable Model of Craniofacial Shape and Texture Variation," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, 2017, pp. 3104–3112. doi:10.1109/ICCV.2017.335

[116] S. Ploumpis, H. Wang, N. Pears, W. A. P. Smith, and S. Zafeiriou, "Combining 3D Morphable Models: A Large Scale Face-And-Head Model," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, pp. 10 926–10 935. doi:10.1109/CVPR.2019.01119

[117] S. Ploumpis, E. Ververas, E. O. Sullivan, et al., "Towards a Complete 3D Morphable Model of the Human Head," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 4142–4160, 2021. doi:10.1109/TPAMI.2020.2991150

[118] B. Egger, W. A. P. Smith, A. Tewari, et al., "3D Morphable Face Models—Past, Present, and Future," *ACM Transactions on Graphics*, vol. 39, pp. 1–38, 2020. doi:10.1145/3395208

[119] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981. doi:10.1145/358669.358692

[120] B. Amberg, S. Romdhani, and T. Vetter, "Optimal Step Nonrigid ICP Algorithms for Surface Registration," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, MN, USA: IEEE, 2007, pp. 1–8. doi:10.1109/CVPR.2007.383165

[121] K. Pearson, "LIII. *On lines and planes of closest fit to systems of points in space*," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, pp. 559–572, 1901. doi:10.1080/14786440109462720

[122] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933. doi:10.1037/h0071325

[123] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004. doi:10.1023/B:VISI.0000013087.49260.fb

[124] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH: IEEE, 2014, pp. 1867–1874. doi:10.1109/CVPR.2014.241

[125] S. Ling, "Generalized Power Method for Generalized Orthogonal Procrustes Problem: Global Convergence and Optimization Landscape Analysis," 2021. doi:10.48550/arXiv.2106.15493

[126] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: A survey," *Medical Image Analysis*, vol. 1, pp. 91–108, 1996. doi:10.1016/S1361-8415(96)80007-7

[127] A. Myronenko and Xubo Song, "Point Set Registration: Coherent Point Drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 2262–2275, 2010. doi:10.1109/TPAMI.2010.46

[128] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–256, 1992. doi:10.1109/34.121791

[129] L. Liang, M. Wei, A. Szymczak, et al., "Nonrigid iterative closest points for registration of 3D biomedical surfaces," *Optics and Lasers in Engineering*, vol. 100, pp. 141–154, 2018. doi:10.1016/j.optlaseng.2017.08.005

[130] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: Reconstruction and parameterization from range scans," *ACM Transactions on Graphics*, vol. 22, pp. 587–594, 2003. doi:10.1145/882262.882311

[131] M. Lüthi, T. Albrecht, and T. Vetter, "Building Shape Models from Lousy Data," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 5762, pp. 1–8. doi:10.1007/978-3-642-04271-3_1

[132] M. A. Styner, K. T. Rajamani, L.-P. Nolte, et al., "Evaluation of 3D Correspondence Methods for Model Building," in *Information Processing in Medical Imaging*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, vol. 2732, pp. 63–75. doi:10.1007/978-3-540-45087-0_6

[133] G. R. Swennen, F. Schutyser, and J.-E. Hausamen, *Three-Dimensional Cephalometry*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. doi:10.1007/3-540-29011-7

[134] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, et al., "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magnetic Resonance Imaging*, vol. 30, pp. 1323–1341, 2012. doi:10.1016/j.mri.2012.05.001

[135] J. L. Hintze and R. D. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism," *The American Statistician*, vol. 52, p. 181, 1998. doi:10.2307/2685478

[136] M. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, p. 3021, 2021. doi:10.21105/joss.03021

[137] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science & Engineering*, vol. 9, pp. 90–95, 2007. doi:10.1109/MCSE.2007.55

[138] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia, "MeshLab: An Open-Source Mesh Processing Tool," *Eurographics Italian Chapter Conference*, p. 8 pages, 2008. doi:10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136

[139] N. Pietroni, M. Tarini, and P. Cignoni, "Almost Isometric Mesh Parameterization through Abstract Domains," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 621–635, 2010. doi:10.1109/TVCG.2009.96

[140] M. Schaufelberger, R. Kühle, A. Wachter, et al., "A Radiation-Free Classification Pipeline for Craniosynostosis Using Statistical Shape Modeling," *Diagnostics*, vol. 12, p. 1516, 2022. doi:10.3390/diagnostics12071516

[141] M. Schaufelberger, R. Kühle, A. Wachter, et al., "A statistical shape model of craniosynostosis patients and 100 model instances of each pathology," 2021. doi:10.5281/ZENODO.6390158

[142] N. Rodriguez-Florez, J. L. Bruse, A. Borghi, et al., "Statistical shape modelling to aid surgical planning: Associations between surgical parameters and head shapes following spring-assisted cranioplasty," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, pp. 1739–1749, 2017. doi:10.1007/s11548-017-1614-5

[143] S. A. H. Tabatabaei, P. Fischer, S. Wattendorf, et al., "Automatic detection and monitoring of abnormal skull shape in children with deformational plagiocephaly using deep learning," *Scientific Reports*, vol. 11, p. 17970, 2021. doi:10.1038/s41598-021-96821-7

[144] R. H. Davies, C. J. Twining, P. Daniel Allen, T. F. Cootes, and C. J. Taylor, "Building optimal 2D statistical shape models," *Image and Vision Computing*, vol. 21, pp. 1171–1182, 2003. doi:10.1016/j.imavis.2003.09.003

[145] H. Lamecker, S. Zachow, H.-C. Hege, M. Zöckler, and E. Haberl, "Surgical treatment of craniosynostosis based on a statistical 3D-Shape model: First clinical application," *International Journal of Computer Assisted Radiology and Surgery*, vol. 1, 2006.

[146] L. M. Abernethy, D. L. England, C. A. Price, P. M. Stevens, and S. R. Wurdeman, "Modified Cephalic Index Measured at Superior Levels of the Cranium Revealed Improved Correction With Helmet Therapy for Patients With Sagittal Suture Craniosynostosis," *Journal of Craniofacial Surgery*, vol. 33, pp. e88–e92, 2022. doi:10.1097/SCS.0000000000008070

[147] S. Ruiz-Correa, R. W. Sze, J. R. Starr, et al., "New Scaphocephaly Severity Indices of Sagittal Craniosynostosis: A Comparative Study with Cranial Index Quantifications," *The Cleft Palate-Craniofacial Journal*, vol. 43, pp. 211–221, 2006. doi:10.1597/04-208.1

[148] W. Schroeder, K. Martin, and B. Lorensen, *The visualization toolkit: An object-oriented approach to 3D graphics ; visualize data in 3D - medical, engineering or scientific ; build your own applications with C++, Tcl, Java or Python ; includes source code for VTK (supports Unix, Windows and Mac)*, 4th ed. Clifton Park, NY: Kitware, Inc, 2006.

[149] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774.

[150] L. Argenta, "Clinical Classification of Positional Plagiocephaly," *Journal of Craniofacial Surgery*, vol. 15, pp. 368–372, 2004. doi:10.1097/00001665-200405000-00004

[151] L. G. Branch, K. Kesty, E. Krebs, L. Wright, S. Leger, and L. R. David, "Argenta Clinical Classification of Deformational Plagiocephaly," *Journal of Craniofacial Surgery*, vol. 26, pp. 606–610, 2015. doi:10.1097/SCS.0000000000001511

[152] K.-k. Shen, J. Fripp, F. Mériaudeau, G. Chételat, O. Salvado, and P. Bourgeat, "Detecting global and local hippocampal shape changes in Alzheimer's disease using statistical shape models," *NeuroImage*, vol. 59, pp. 2155–2166, 2012. doi:10.1016/j.neuroimage.2011.10.014

[153] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[154] M. Schaufelberger, C. Kaiser, R. Kühle, et al., "3D-2D Distance Maps Conversion Enhances Classification of Craniosynostosis," *IEEE Transactions on Biomedical Engineering*, pp. 1–10, 2023. doi:10.1109/TBME.2023.3278030

[155] J. d'Errico, "Matlab central file exchange: inpaint_nans," 2022, https://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint_nans, retrieved 2022-07-15.

[156] Maisevector, "KIT-IBT/cd-map: Initial release of Craniosynostosis Distance Maps (CD-Map)," Zenodo, 2022. doi:10.5281/ZENODO.7253975

[157] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015. doi:10.48550/arXiv.1512.03385

[158] C. Szegedy, W. Liu, Y. Jia, et al., "Going Deeper with Convolutions," 2014. doi:10.48550/arXiv.1409.4842

[159] A. G. Howard, M. Zhu, B. Chen, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017. doi:10.48550/arXiv.1704.04861

[160] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," 2017. doi:10.48550/ARXIV.1703.01365

[161] N. Kokhlikyan, V. Miglani, M. Martin, et al., "Captum: A unified and generic model interpretability library for PyTorch," 2020. doi:10.48550/arXiv.2009.07896

[162] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL: IEEE, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848

[163] P.-J. Kindermans, S. Hooker, J. Adebayo, et al., "The (Un)reliability of Saliency Methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learn-*

ing. Cham: Springer International Publishing, 2019, vol. 11700, pp. 267–280. doi: 10.1007/978-3-030-28954-6_14

[164] C. Nagel, M. Schaufelberger, O. Dössel, and A. Loewe, "A Bi-atrial Statistical Shape Model as a Basis to Classify Left Atrial Enlargement from Simulated and Clinical 12-Lead ECGs," in *Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge*. Cham: Springer International Publishing, 2022, vol. 13131, pp. 38–47. doi: 10.1007/978-3-030-93722-5_5

[165] J. Sánchez, G. Luongo, M. Nothstein, et al., "Using Machine Learning to Characterize Atrial Fibrotic Substrate From Intracardiac Signals With a Hybrid in silico and in vivo Dataset," *Frontiers in Physiology*, vol. 12, p. 699291, 2021. doi:10.3389/fphys.2021.699291

[166] T. Pinetz, J. Ruisz, and D. Soukup, "Actual Impact of GAN Augmentation on CNN Classification Performance," in *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*. Prague, Czech Republic: SCITEPRESS - Science and Technology Publications, 2019, pp. 15–23. doi:10.5220/0007244600150023

[167] M. Arjovsky and L. Bottou, "Towards Principled Methods for Training Generative Adversarial Networks," 2017. doi:10.48550/arXiv.1701.04862

[168] C. Kaiser, M. Schaufelberger, R. Kühle, et al., "Generative-Adversarial-Network-Based Data Augmentation for the Classification of Craniosynostosis," *Current Directions in Biomedical Engineering*, vol. 8, pp. 17–20, 2022. doi:10.1515/cdbme-2022-1005

[169] M. Schaufelberger, R. Kühle, A. Wachter, et al., "GAN, PCA, and Statistical Shape Models for the Creaiton of Synthetic Craniosynostosis Distance Maps," 2023. doi: 10.5281/ZENODO.8117499

[170] W. Bai, W. Shi, A. de Marvao, et al., "A bi-ventricular cardiac atlas built from 1000+ high resolution MR images of healthy subjects and an analysis of shape and motion," *Medical Image Analysis*, vol. 26, pp. 133–145, 2015. doi:10.1016/j.media.2015.08.009

[171] T.-N. Nguyen, V.-D. Tran, H.-Q. Nguyen, and T.-T. Dao, "A statistical shape modeling approach for predicting subject-specific human skull from head surface," *Medical & Biological Engineering & Computing*, vol. 58, pp. 2355–2373, 2020. doi: 10.1007/s11517-020-02219-4

[172] N. Chakravorty, C. S. Sharma, K. A. Molla, and J. K. Pattanaik, "Open Science: Challenges, Possible Solutions and the Way Forward," *Proceedings of the Indian National Science Academy*, vol. 88, pp. 456–471, 2022. doi:10.1007/s43538-022-00104-2

[173] M. W. Jeter, *Mathematical programming: An introduction to optimization*, Monographs and Textbooks in Pure and Applied Mathematics, no. 102. New York: M. Dekker, 1986.

List of Publications and Supervised Theses

Journal Articles

- **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jürgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *A Radiation-Free Classification Pipeline for Craniosynostosis Using Statistical Shape Modeling*, Diagnostics 2022, 12(7), 1516
- **Matthias Schaufelberger**, Christian Kaiser, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jörgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *3D-2D Distance Maps Conversion Enhances Classification of Craniosynostosis*, IEEE Transactions on Biomedical Engineering, vol. 70, no. 11, pp. 3156–3165, 2023
- **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jürgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *Impact of Data Synthesis Strategies for the Classification of Craniosynostosis*, Frontiers in Medical Technology, 5:1254690, 2023
- Reinald Kühle, Friedemann Ringwald, Frederic Bouffleur, Niclas Hagen, **Matthias Schaufelberger**, Werner Nahm, Jürgen Hoffmann, Christian Freudlsperger, Michael Engel, Urs Eisenmann *The Use of Artificial Intelligence for the Classification of Craniofacial Deformities*, Journal of Clinical Medicine, Journal of Clinical Medicine 2023, 12(22), 7082
- Robin Andlauer, Andreas Wachter, **Matthias Schaufelberger**, Frederic Weichel, Reinald Kühle, Christian Freudlsperger, Werner Nahm *3D-Guided Face Manipulation of 2D Images for the Prediction of Post-Operative Outcome After Cranio-Maxillofacial Surgery*, IEEE Transactions on Image Processing, vol. 30, pp. 7349–7363, 2021
- Steffen Schuler, **Matthias Schaufelberger**, Laura R. Bear, Jake A. Bergquist, Matthijs J. M. Cluitmans, Jaume Coll-Font, Önder N. Onak, Brian Zenger, Axel Loewe, Rob S. MacLeod, Dana H. Brooks, Olaf Dössel *Reducing Line-of-Block Ar-*

tifacts in Cardiac Activation Maps Estimated Using ECG Imaging: A Comparison of Source Models and Estimation Methods, vol. 69, no. 6, pp. 2041–2052

Refereed Conference Articles

- **Matthias Schaufelberger**, Reinald Kühle, Frederic Weichel, Andreas Wachter, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Christian Freudlsperger, Werner Nahm *Laplace-Beltrami Refined Shape Regression Applied to Neck Reconstruction for Craniosynostosis Patients*, Current Directions in Biomedical Engineering, vol. 7(2), pp. 191–194, 2021
- **Matthias Schaufelberger**, Reinald Kühle, Christian Kaiser, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Christian Freudlsperger, Werner Nahm *CNN-Based Classification of Craniosynostosis Using 2D Distance Maps*, 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 446–449
- Christian Kaiser, **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Michael Engel, Christian Freudlsperger, Werner Nahm *Generative-Adversarial-Network-Based Data Augmentation for the Classification of Craniosynostosis*, Current Directions in Biomedical Engineering, vol. 8(2), pp. 17–20, 2022
- Anna Maria Becker, **Matthias Schaufelberger**, Reinald Kühle, Christian Freudlsperger, Werner Nahm *Multi-Height Extraction of Clinical Parameters Improves Classification of Craniosynostosis*, Current Directions in Biomedical Engineering, vol. 9, in press, 2023
- **Matthias Schaufelberger**, Steffen Schuler, Laura R. Bear, Matthijs J. M. Cluitmans, Jaume Coll-Font, Önder N. Onak, Olaf Dössel, Dana H. Brooks *Comparison of Activation Times Estimation for Potential-Based ECG Imaging*, Computing in Cardiology, vol. 46, 2019
- Claudia Nagel, **Matthias Schaufelberger**, Olaf Dössel, Axel Loewe, *A Bi-atrial Statistical Shape Model as a Basis to Classify Left Atrial Enlargement from Simulated and Clinical 12-Lead ECGs*, Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge, 2021, pp. 38–47
- Denis Krnjaca, Lorena Krames, **Matthias Schaufelberger**, Werner Nahm *A Statistical Shape Model Pipeline to Enable the Creation of Synthetic 3D Liver Data*, Current Directions in Biomedical Engineering, vol. 9, in press, 2023

Refereed Conference Abstracts

- **Matthias Schaufelberger**, Reinald Kühle, Frederic Weichel, Christian Freudlsperger, Werner Nahm *Testing a Point Distribution Model of the Head Designed From Healthy Subjects in Respect of Craniofacial Deformities*, 54th Annual Conference of the German Society of Biomedical Engineering, Leipzig (online), 2020
- **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Christian Freudlsperger, Michael Engel, Werner Nahm *A Publicly Available Statistical Shape Model for the Assessment of Craniosynostosis*, 2022 Joint Annual Conference of the Austrian (ÖGBMT), German (VDE DGBMT) and Swiss (SSBE) Societies for Biomedical Engineering, Innsbruck, 2022

Conference Presentations

- **Matthias Schaufelberger**, Reinald Kühle, Frederic Weichel, Christian Freudlsperger, Werner Nahm *Testing a Point Distribution Model of the Head Designed From Healthy Subjects in Respect of Craniofacial Deformities*, 54th Annual Conference of the German Society of Biomedical Engineering, Leipzig (scheduled poster presentation, online), 2020
- **Matthias Schaufelberger**, Reinald Kühle, Christian Kaiser, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Christian Freudlsperger, Werner Nahm *CNN-Based Classification of Craniosynostosis Using 2D Distance Maps*, 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow (scheduled poster presentation), 2022

Talks

- KCIST Online Lecture Series zu KI: Classification of Head Deformities Using Statistical Shape Modeling, 07.12.2020, Karlsruhe (online)

Awards

- Student competition finalist, 57th DGBMT Annual Conference on Biomedical Engineering, Duisburg, Germany, 2023, Anna Maria Becker, **Matthias Schaufelberger**, Reinald Kühle, Christian Freudlsperger, Werner Nahm, *Multi-Height Extraction of Clinical Parameters Improves Classification of Craniosynostosis*

Theses and Student Research Projects

- *Activation Times Estimation in ECG Imaging: Comparison of Source Models and Estimation Methods*, Master's Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2019
- *Deep-learning supported gait analysis*, Student Research Project, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2019

Code and Data Repositories

- **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jürgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *A statistical shape model of craniosynostosis patients and 100 model instances of each pathology*, Zenodo, `https://zenodo.org/record/5638148`
- **Matthias Schaufelberger**, Christian Kaiser, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jörgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *Craniosynostosis Distance Maps (CD-Maps)*, `https://github.com/KIT-IBT/cd-map`
- **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jürgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *GAN, PCA, and Statistical Shape Models for the Creation of Synthetic Craniosynostosis Distance Maps*, Zenodo, `https://zenodo.org/record/8117499`
- **Matthias Schaufelberger**, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jörgen Hoffmann, Michael Engel, Christian Freudlsperger, Werner Nahm *Craniosource-GAN-PCA-SSM*, `https://github.com/KIT-IBT/craniosource-gan-pca-ssm`

Supervised Students

- Hasan Bahadır Savaş, *Classification of Craniosynostosis on 2D Distance Maps Using Convolutional Neural Networks*, Bachelor's Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2021
- Christian Kaiser, *Ray-based Assessment of Craniosynostosis: Classification, Data Augmentation and Feature Analysis*, Master's Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2022
- Samir Schürmann, *Shape-Model-Based Skull Estimation From Head Surface*, Master's Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2022
- Denis Krnjaca, *Generation of a Statistical Shape Model for the Liver*, Bachelor's Thesis (Co-Supervision), Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2022
- Anna Maria Becker, *Evaluation and Optimization of the Decision Tree Model for the Classification of Head Deformities in Clinical Practice*, Bachelor's Thesis, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2023
- Sudhanshu Shrivastava, *Shape-Model-Based Skull Estimation of Head Deformities*, student research project, Institute of Biomedical Engineering, Karlsruhe Institute of Technology (KIT), 2023