3D-2D Distance Maps Conversion Enhances Classification of Craniosynostosis

Matthias Schaufelberger, Christian Kaiser, Reinald Kühle, Andreas Wachter, Frederic Weichel, Niclas Hagen, Friedemann Ringwald, Urs Eisenmann, Jürgen Hoffmann, Michael Engel, Christian Freudlsperger, and Werner Nahm

Abstract—**Objective:** Diagnosis of craniosynostosis using photogrammetric 3D surface scans is a promising radiation-free alternative to traditional computed tomography. We propose a 3D surface scan to 2D distance map conversion enabling the usage of the first convolutional neural networks (CNNs)-based classification of craniosynostosis. Benefits of using 2D images include preserving patient anonymity, enabling data augmentation during training, and a strong under-sampling of the 3D surface with good classification performance. **Methods:** The proposed distance maps sample 2D images from 3D surface scans using a coordinate transformation, ray casting, and distance extraction. We introduce a CNN-based classification pipeline and compare our classifier to alternative approaches on a dataset of 496 patients. We investigate into low-resolution sampling, data augmentation, and attribution mapping. **Results:** Resnet18 outperformed alternative classifiers on our dataset with an F1-score of 0.964 and an accuracy of 98.4%. Data augmentation on 2D distance maps increased performance for all classifiers. Under-sampling allowed 256-fold computation reduction during ray casting while retaining an F1-score of 0.92. Attribution maps showed high amplitudes on the frontal head. **Conclusion:** We demonstrated a versatile mapping approach to extract a 2D distance map from the 3D head geometry increasing classification performance, enabling data augmentation during training on 2D distance maps, and the usage of CNNs. We found that low-resolution images were sufficient for a good classification performance. **Significance:** Photogrammetric surface scans are a suitable craniosynostosis diagnosis tool for clinical practice. Domain transfer to computed tomography seems likely and can further contribute to reducing ionizing radiation exposure for infants.

Matthias Schaufelberger is with the Institute of Biomedical Engineering, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (e-mail: publications@ibt.kit.edu).

Christian Kaiser, Andreas Wachter, and Werner Nahm are with the Institute of Biomedical Engineering, Karlsruhe Institute of Technology, Germany.

Reinald Kühle, Frederic Weichel, Jürgen Hoffmann, Michael Engel, and Christian Freudlsperger are with the Department of Oral and Maxillofacial Surgery, Heidelberg University Hospital, Germany.

Niclas Hagen, Friedemann Ringwald, and Urs Eisenmann are with the Institute of Medical Informatics, Heidelberg University Hospital, Germany.

I. Introduction

A. Craniosynostosis

CRANIOSYNOSTOSIS is characterized by the premature fusion of skull sutures resulting in irregular growth patterns. It affects infants and its prevalence is estimated to be three to six cases per 10,000 live births [1], [2]. For isolated cases craniosynostosis is distinguished into different subtypes, depending on the prematurely fused suture: sagittal synostosis (scaphocephaly), metopic synostosis (trigonocephaly), unilateral coronal synostosis (anterior plagiocephaly), lambda synostosis (posterior plagiocephaly), and bicoronal synostosis (brachycephaly). Although brachycephaly synostosis includes the fusion of both coronal sutures, the medical community counts it among isolated synostosis. As dictated by Virchow's law, the premature closure of a suture blocks the expansion perpendicular to the suture and causes compensatory growth parallel to it, leading to distinct head deformities for each type [3]. Craniosynostosis can lead to elevated intracranial pressure [4] which is linked to reduced brain growth and diminished neuropsychological development [5]. To decrease the intracranial pressure and enable a normal skull development, surgical remodeling of the skull is the most common therapy. Complications are rare [6] and in most cases a normalized head shape can be achieved [7]. For further reading about craniosynostosis, the reader is referred to [8].

Early diagnosis is crucial and can only be performed in specialized hospitals. During diagnosis, a combination of visual examination, palpation, cephalometric measurements, and medical imaging is performed. Traditional computed tomography (CT) imaging is considered the gold standard, but exposes the infants to ionizing radiation [9]. Black bone magnetic resonance imaging (MRI) [10] avoids the harmful radiation but requires sedation or general anesthesia to prevent the children from moving. Sonographic examinations [11] and 3D photogrammetry are radiation-free alternatives. Especially photogrammetric 3D scans provide inexpensive and fast means to objectively quantify head shape suitable for monitoring head development before and after surgery [12]. A reliable and fast, machine-aided classification using 3D surface scans might eliminate the use
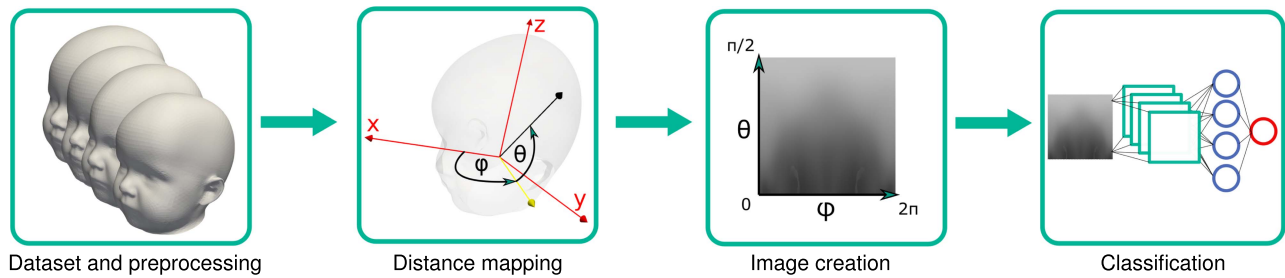
Fig. 1. 2D distance map classification pipeline. Each dataset sample is preprocessed to remove corruptions. After distance extraction according to the mapping approach, the image can be assembled, and a CNN-based classification can be performed.

of ionizing radiation during CT-based diagnosis and could be performed by pediatricians within minutes.

B. Classification of Craniosynostosis

Automatic classification approaches of craniosynostosis have been proposed on CT data, 3D photogrammetric surface scans, and 2D images from different perspectives. Early CT-based studies were performed on 2D data [13] to distinguish between scaphocephaly and a control group. Approaches on 3D data used a combination of crafted features and statistical modeling [14]. Recently, the use of radiation-free 3D stereophotographs for head shape assessment and classification of craniosynostosis gained momentum. In our own group we developed a statistical shape model (SSM)-based classification approach [15]. Ray-based classification of craniosynostosis was proposed by [16], who extracted distances from a defined center point using a feedforward neural network (FNN) with an accuracy of 99.5%. Image and machine-learning-based classification approaches for the classification of plagiocephaly were proposed [17], but were not applied to craniosynostosis. Multi-image 2D approaches could identify different types of craniosynostosis with feature-based classifiers and a pre-trained CNN model [18].

CNNs are popular image-based classifiers and often a good choice for classification problems due to a flexible model design, the easy adaption of many pre-trained models facilitating transfer learning, and the possibility to perform many different types of data augmentation during training. However, for classifying 3D data, a transformation from 3D to 2D is required, which had not been proposed for classifying head deformities. Data augmentation on the 3D data as in [16] was therefore limited to adding noise to the input features. 3D transformations such as left-right patient mirroring and rotational misalignment have to be applied before training and cannot be randomized during training as the distance extraction is computationally expensive.

C. Scope of This Work

The first contribution of this work is a mapping approach to obtain 2D images from 3D surface scans of the head. We combined ideas from [19] who proposed asymmetry maps for plagiocephaly patients and [16] who used ray casting for distance extraction. 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.

Regarding classification, the encoding of the 3D geometry onto a 2D image enables using CNNs on the image domain. 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. Also, 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 we show that classification performance can be maintained while systematically reducing image resolution. The second contribution of this work is the first CNN-based classification of craniosynostosis using 3D surface scans, which is enabled by the usage of the 2D distance maps. Introducing this new classification approach, we take the opportunity to also conduct the first comparison study of craniosynostosis classifiers on, to the best of our knowledge, the largest dataset of craniosynostosis patients to date. We consider two alternative approaches [15], [16] for the first time tested on the same patients, enabling a quantitative comparison of different classifiers under the same conditions. For the benefit of the community, we released the Python modules for the distance map creation which can also be used on our previously published SSM [20].

II. Materials and Methods

Fig. 1 gives a full overview of the pipeline from the raw data to the distance map creation and the craniosynostosis classification.

A. Dataset and Preprocessing

We obtained the photogrammetric surface scans from the Department of Oral and Maxillofacial Surgery of the Heidelberg University Hospital, where patients with craniofacial diseases
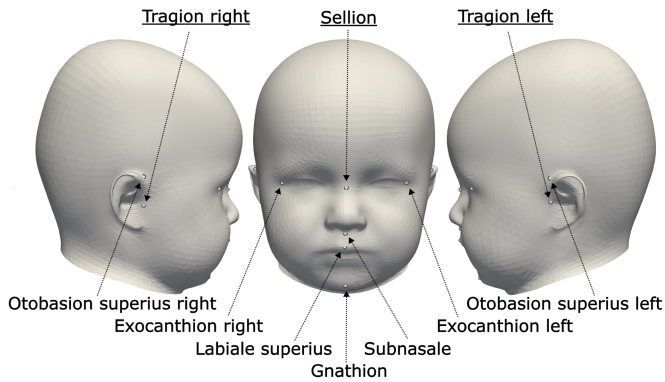
Fig. 2. Landmarks provided in the dataset. The chosen landmarks for the coordinate system were sellion, and left and right tragion (underlined and on top).
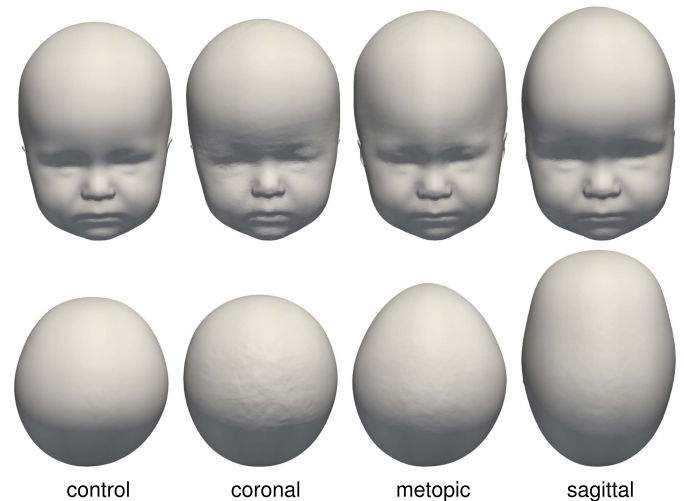

Fig. 3. Head shapes of the four classes in the dataset. Top row: front view, bottom row: top view.
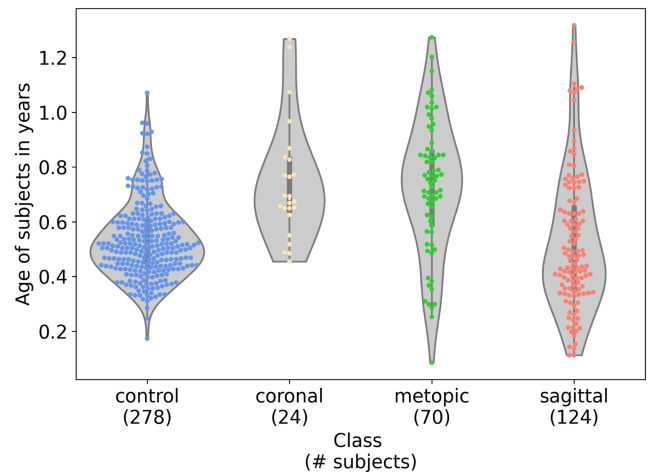

Fig. 4. Top: The type of head deformities for the four classes. Bottom: Class and age distribution of the subjects in the dataset. Parenthesis indicate number of samples per class.

are routinely recorded for monitoring and documentation purposes. We used a standardized protocol, 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 parents. The 3D surface scans were recorded using a 3D image recording system (Canfield VECTRA-360-nine-pod, Canfield Science, Fairfield, NJ, USA) and the children wore tight-fitting hairnet caps to minimize artifacts caused by the hair. For each recording, we obtained a triangular surface mesh which was later annotated with 10 cephalometric landmarks and the medical diagnosis by the surgeon. The available landmarks are visualized in Fig. 2.

Out of the scans that were acquired between 2011 and 2021, we extracted only the preoperative scans closest to the operation date to avoid duplicate scans of the same patient. We selected craniosynostosis patients with coronal (brachycephaly and unilateral anterior plagiocephaly), sagittal (scaphocephaly), or metopic suture fusion (trigonocephaly), as well as a control group without any suture fusion. Besides healthy subjects, our control group also contained scans of children with mild positional posterior plagiocephaly. While positional plagiocephaly patients were later treated with helmet therapy or laying repositioning, all craniosynostosis patients underwent surgical remodeling of the cranium. The four classes are displayed in Fig. 3. The final dataset consisted of 496 subjects. A violin plot [21], [22], [23] of the 496 patients' class and age distribution is displayed in Fig. 4. Regarding the selection of classes, our approach is comparable to other classification studies, which distinguished between craniosynostosis and non-craniosynostosis classes, in particular [14], [15], [16].

We used the Python module pymeshlab from the open-source software Meshlab [24] to preprocess the 3D surface scans. We removed isolated parts, duplicate faces and vertices, and closed holes in the surface scans in a fully automated manner, as those types of artifacts could lead to incorrect data in the distance maps. Additionally, parts at the ears were often characterized by large edge lengths, so we used isotropic explicit remeshing [25] with a target length of 1 mm 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. After alignment using the landmark-based coordinate system, everything below the child's neck can be cut off to speed up computation during image creation.

B. Distance Mapping

We used the patients' anatomic landmarks for the creation of a common coordinate system similar to the frontal, sagittal, and median planes. For a coordinate system we chose left and right tragion (located on the ears), as well as the sellion (located on the nose), as they were located on different ends of the head and because the midpoint between the two ears is approximately in the center of the head. The location of the center point is similar to the computed cranial focus point definition [26] for CT data. We defined the center point $\mathbf{p}_c$ as the midpoint of left and right tragion ($\mathbf{p}_{tl}$ and $\mathbf{p}_{tr}$) to serve as the origin of the new coordinate

frame:

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

The axis direction $\mathbf{u}_x$ was defined as the frontal axis with the direction from the origin to the sellion located on the nose:

$$\mathbf{u}_x = \mathbf{p}_s - \mathbf{p}_c \tag{2}$$

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

$$\mathbf{u}_y = (\mathbf{p}_{tl} - \mathbf{p}_c) - \mathbf{u}_x \frac{\mathbf{u}_x \cdot (\mathbf{p}_{tl} - \mathbf{p}_c)}{\|\mathbf{u}_x\|}. \tag{3}$$

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

$$\mathbf{u}_z = \mathbf{u}_x \times \mathbf{u}_y \tag{4}$$

The direction vectors $[\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z]^\mathsf{T}$ were each normalized to an orthonormal basis $[\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z]^\mathsf{T}$.

Two angle directions to sample the head surface for the 2D image were set up. We proposed three different mapping approaches each requiring a specific coordinate frame to map the points from the Cartesian space to the image domain: Spherical, arch-spherical, and cylindrical, visualized in Fig. 5. The mathematical descriptions of the three mappings are explained below.

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

$$\mathbf{d}_s = \left[\cos\varphi\cos\theta, \sin\varphi\cos\theta, \sin\theta\right]^\mathsf{T}. \tag{5}$$

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

$$\mathbf{p}_s = \mathbf{p}_c \tag{6}$$

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, we placed the image origin in the bottom left corner and defined the direction for θ from the lower left corner to the upper left corner (see Fig. 5, top right panel).

2) Arch-Spherical: 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 $\mathbf{d_a}$

$$\mathbf{d}_a = \left[\sin\varphi\cos\theta, \cos\varphi, \sin\varphi\sin\theta\right]^\mathsf{T}, \tag{7}$$

and the start point again being placed in the origin:

$$\mathbf{p}_s = \mathbf{p}_c \tag{8}$$

The corresponding angle intervals for φ and θ both ranged from $0 \leq \varphi < \pi$, but the direction in which we measured φ was defined clockwise (mathematically negative) to retain the left-right relation.

3) Cylindrical: For the two spherical-based mappings, a grayscale gradient could be observed from the larger height than width of the human head. By using a cylindrical-based mapping instead of a spherical one, we could reduce the distance variation
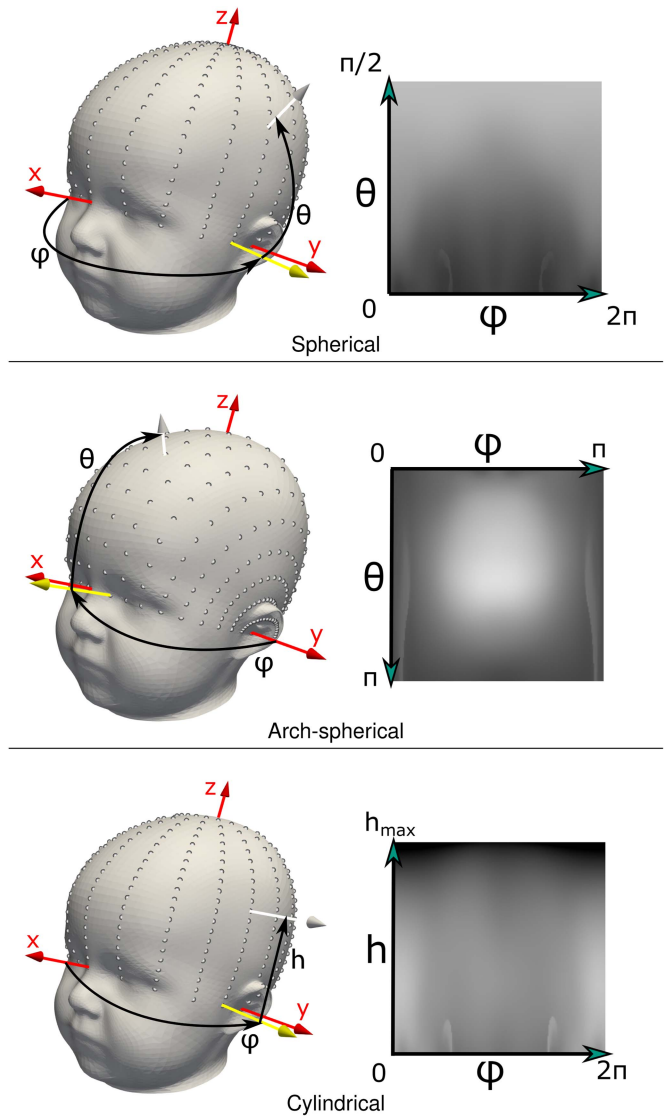

Fig. 5. Visualization of the mapping types. Left: Angle definitions and coordinate frames. Hit points from the rays resulting from a 20 × 20 sampling are visualized. Right: Distance maps corresponding from the mapping with angle axes.

for a larger part of the image. One key feature was that the center point was not constant, but moved towards the tip of the head for each pixel row. Thus, the direction $\mathbf{d_c}$ was defined as

$$\mathbf{d}_c = \left[\cos\varphi, \sin\varphi, 0\right]^\mathsf{T}, \tag{9}$$

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

$$\mathbf{p}_s = \mathbf{p}_c + h \cdot \mathbf{u}_z. \tag{10}$$

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

C. Image Creation

Each angle interval was sampled equidistantly in 224 steps, resulting in one ray direction for each of the 2D image pixels. We determined the intersection of the 3D mesh surface with each ray and extracted the distance from the starting point to the hit point. We used oriented bounding boxes (OBB) trees from the vtk Python package [27] 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. 5). If multiple hit points were encountered (for example at the auricula, the outermost part of the ear), we chose the minimal distance as the "correct" distance. If no hit point could be determined (for example due to corruptions in the scans), we interpolated missing values on the equally-spaced image grid as the mean of its four neighbors (which models missing pixels according to Laplace's equation) [28]. Small artifacts resulted from the tip of the head which we left unchanged. However, if required, they might be minimized on the image domain using smoothing. We created the actual image by converting the distances to integer pixel intensity values from 0 to 255. We explored two different normalization schemes to transform the distance range to the required image intensity range: Linear re-scaling and per-pixel-based re-scaling.

Linear re-scaling uses only *one* linear transformation for *all images and pixel* values. We computed mean distance and standard deviation 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, 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 direction). This way, the intensity range is better sampled for each pixel, but 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. 6 shows the different scaling approaches for the same distance map and Fig. 7 the different mapping types. Other normalization approaches might also be possible, e.g., scaling with respect to a control group or min-max scaling irrespective to other subjects. Our Python source code[1] [29] for the distance map creation was made publicly available and can be combined with our synthetic dataset [20].

D. 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 our proposed method.
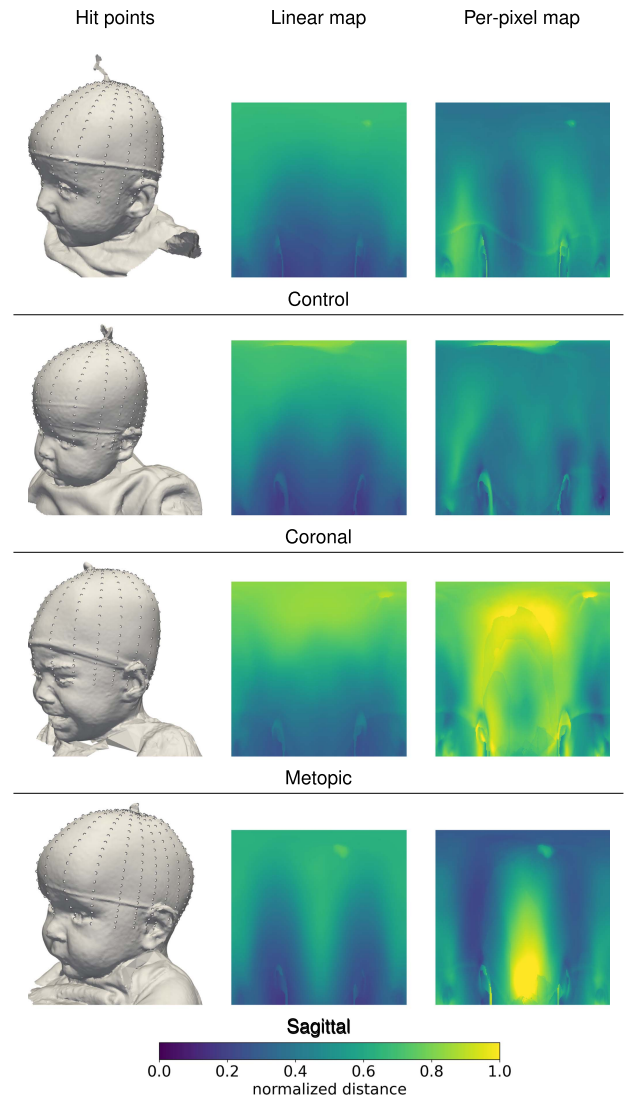
Fig. 6. Linear and per-pixel scaling applied to the different pathologies using the spherical mapping. For the visualization of the hit points, we used 20 × 20 rays instead of 224 × 224. For visualization purposes we used a blue-yellow colormap instead of grayscale.

1) Classification Comparison: The first test was designed as a comparison between different CNNs using our proposed distance maps and alternative approaches from the literature.

We considered a vanilla CNN trained from scratch only on the image data. To find the optimal number of convolutional layers, we trained the CNN 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 we further considered. For the pre-trained CNNs, this included Resnet18 [30], AlexNet [31], GoogLeNet [32], and small and large Mobilenet [33].

As alternative approaches we considered SSM-based classification using linear discriminant analysis (LDA) [15] and FNN-based classification [16]. The SSM-based approach was developed in our own group, so we trained the classifier on the updated dataset. Regarding the FNN classifier [16], we re-implemented the FNN as described in the original paper [16].
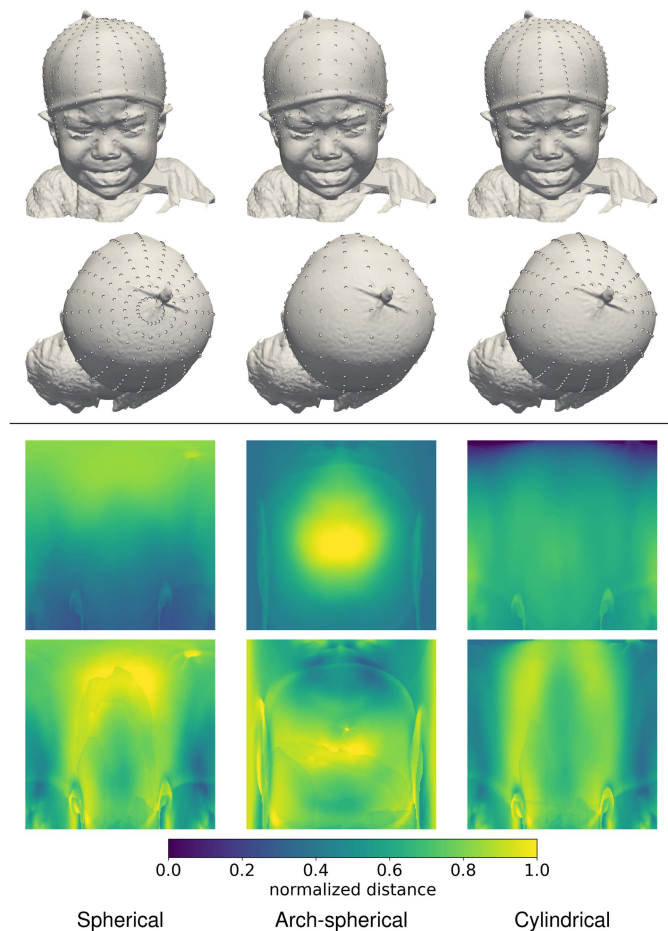
Fig. 8. True 224 × 224 image in comparison with the three different interpolation methods nearest neighbors, bilinear, and bicubic from a 7 × 7 image.

Fig. 7. Mapping methods with different scalings for one subject. For the visualization of the hit points, we used 20 × 20 rays 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 we used a blue-yellow colormap instead of grayscale.

After initial testing, we removed the dropout and batch-norm-layers which improved the performance on our dataset. The FNN-based approach was tested on an icosphere-based extraction scheme as originally proposed [16] and on the distances extracted using our proposed mapping. This way, the FNN could be tested on both inputs.

2) Mapping Comparison: We compared the three mappings and two scaling approaches using the same network (pre-trained Resnet18) to test if the mapping had a substantial influence on the performance metrics. We show an exemplary set of images derived with the different mappings in Fig. 7.

3) Data Augmentation Strategy: We tested image-based data augmentation on pre-trained Resnet18, the vanilla CNN, and the FNN on the linear, spherical mapping. We included four types of data augmentation: 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 can be interpreted as enlarging or shrinking the full head. Random flipping in horizontal direction
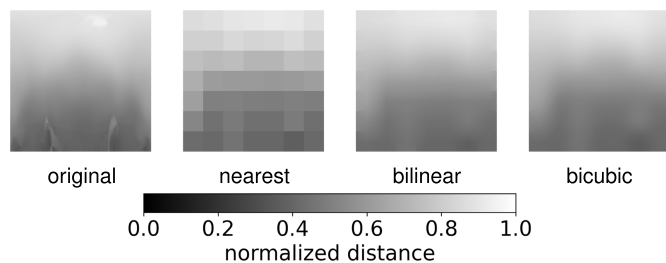
was applied with a probability of $p = 0.5$ and corresponds to a symmetric mirroring of the patients. The horizontal shift was designed as white Gaussian noise with standard deviation of $\sigma = 20/360 \cdot 2\pi$, shifting the image to one direction and inserting the cut-off part on the other side. This corresponds to a head rotation during recording, as if the subjects were looking slightly left or right.

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

4) Resolution: This test 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. We tested to use only $n \times n$ rays with n ranging from 7 to 224 in steps of 7. We interpolated the smaller images with intermediate points to obtain the CNN input dimension of 224. We tested three interpolation methods: nearest-neighbor-mapping, bilinear and bicubic image interpolation (Fig. 8). We used again a pre-trained Resnet18 and the linearly scaled spherical mapping to ensure comparability among the experiments.

5) Attribution Maps: We intended to visualize which parts of the image (and consequently from the 3D surface scans) contributed to the model's predictions by using integrated gradients [34]. 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). We used the captum package [35] for computing integrated gradients. For visualizing the heatmap resulting from the integrated gradients on the 3D head surface, we projected each point of the 3D surface onto the image, bilinearly interpolated the respective attribution value, and back-projected the attribution value to each 3D point. We used the following three different back-transformations for the spherical, arch-spherical, and cylindrical transformation:

$$\begin{bmatrix} r, \varphi, \theta \end{bmatrix}^{\top}_{\text{spherical}} = \begin{bmatrix} \sqrt{x^2 + y^2 + z^2} \\ \arccos(z/r) \\ \text{atan2}(y, x) \end{bmatrix} \quad (11)$$

Classifier	Training time	Accuracy	G-mean	F1-score
Alternative methods from the literature				
LDA on SSM [15]	5 s	0.964±0.027	0.715±0.365	0.890±0.100
FNN [16] on semi-icosphere	130 s	0.954±0.031	0.696±0.355	0.876±0.104
Alternative methods from the literature with distance maps				
FNN [16] 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

The bold values denote the best results.

$$\left[r, \varphi, \theta\right]^{\mathsf{T}}_{\mathrm{arch-spherical}} = \begin{bmatrix} \sqrt{x^2 + y^2 + z^2} \\ \arcsin\left(y/r\right) \\ \mathrm{atan2}(z, x) \end{bmatrix} \quad (12)$$

$$\left[\varphi, \rho, z\right]^{\mathsf{T}}_{\mathrm{cylindrical}} = \begin{bmatrix} \mathrm{atan2}(y, x) \\ \sqrt{x^2 + y^2} \\ z \end{bmatrix} \quad (13)$$

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 neural networks were trained with cross entropy loss, 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 we trained from-scratch networks with 1000 epochs and a step size of 100. The SSM-based classifier was trained with the same hyperparameters as in our previous work [15] on the new dataset. We built our framework on Pytorch [36] and Scikit-learn [37]. Pytorch's pre-trained models were trained on ImageNet [38]. For model evaluation, we used accuracy, g-mean, and macro F1-score and computed mean values and standard deviations across all cross validation splits.

III. RESULTS

A. Classification Comparison

As summarized in Table I, all classifiers showed good performance with mean accuracies above 0.95. CNN-based classifiers generally performed better than the FNN and the SSM. The highest accuracies, g-means, and F1-scores were achieved by GoogLeNet and Resnet18. Standard deviations for F1-score and g-mean computed across the ten folds were lowest for GoogLeNet and Resnet18 (indicating smaller disturbances for

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

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

The bold values denote the best results.

different training conditions) and increased for the other networks. The CNNs scored higher accuracies, g-means, and F1-scores than the alternative FNNs. In general, accuracy ranged from 0.954 to 0.984 which corresponds to 15 fewer misclassified test samples. We provide the confusion matrix with sensitivities and specificities of the pre-trained Resnet18 classifier in Table II.

We included training times for each cross validation split measured on a high performance cluster running Red Hat Enterprise Linux using a Nvidia Tesla V100. 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 can processed in parallel since they are independent of each other. Training times for each classifier are included in Table I.

B. Mapping Comparison

We display classification results for different mapping approaches using the pre-trained Resnet18 in Table III. All accuracies were 0.976 or above. All three metrics were consistently higher than the classification approaches in Table I 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.

C. Data Augmentation Strategy

Table IV shows the classifier performance using "on the fly" 2D image-based data augmentation, compared to the networks without data augmentation (Table I). All classifiers improved g-mean and F1-score. FNN showed the largest improvements in all three metrics.

D. Resolution

In Fig. 9, we show the cross validation mean of accuracy, g-mean, and F1-score over pixel resolution for the Resnet18

Classifier	Accuracy	G-mean	F1-score
FNN [16] on distance maps	0.968±0.024	0.888±0.083	0.928±0.055
(Mean improvement w.r.t. Table I)	**(+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 I)	(+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 I)	(+0.002)	(+0.026)	(+0.011)

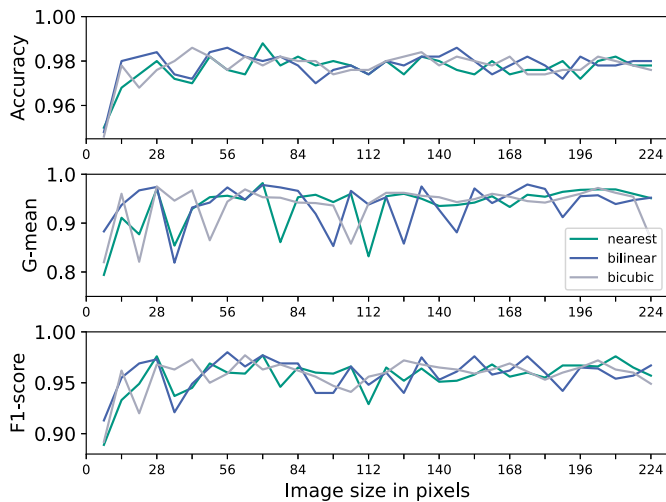The bold values denote the best results.

Fig. 9. Mean cross validation metrics as functions of the number of pixels p to create an $p \times p$ image. Three different interpolation methods were used to create an up-scaled image: Nearest neighbors, bilinear, and bicubic interpolation.

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 rays per direction resulted in a 256-fold computation reduction for the ray tracing while bicubic image interpolation still yielded a g-mean larger than 0.95. All three interpolation methods jittered slightly and with similar amplitudes.

E. Attribution Maps

Fig. 10 indicates 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 and cylindrical mappings, and slightly shifted to the left for the arch-spherical mapping.

IV. DISCUSSION

We proposed a flexible mapping approach to create 2D distance maps from 3D head geometries, which we used for
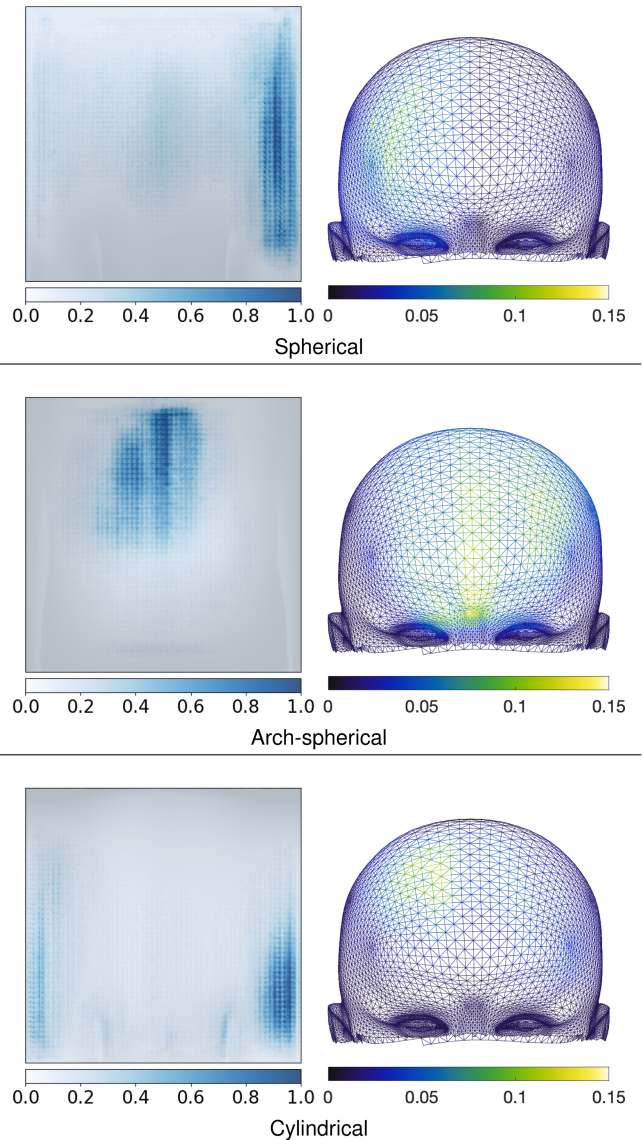

Fig. 10. Mean attribution for all subjects in the image domain with a transparent overlay of the map (left) and projected onto the 3D surface for the respective mapping (right). From top to bottom: Spherical, arch-spherical, and cylindrical mapping. The 2D image shows a transparent overlay of the grayscaled distance map as a visual guide, while attribution is colored in blue. A larger value means higher attribution.

the classification of craniosynostosis. We introduced multiple mapping variants with different coordinate systems and scaling approaches. We extended the ideas of [16] and structured the extracted distances in a 2D grid. While CNNs had been used for camera pictures from above [18], we propose an encoding strategy which include the 3D data encoded in 2D image intensity and employed the first CNN for craniosynostosis on such a type of encoding. 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 and is a suitable option for a cross-domain dataset.

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 improvement 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 rays per direction 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 spatial frequencies well and suppress high spatial frequencies, it suggests that low spatial frequencies 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 towards 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 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 [39].

There are also limitations to this study. As with many studies in biomedical engineering, our 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 patients 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 often complicated due to patient data regulations. Other groups might make use of our publicly available SSM [20]. Data augmentation or data synthesis might be an option to make the classification models as robust as possible. We showed that image-based random horizontal flipping and a random horizontal shift during training improved 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. 7). 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, our 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 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.

V. CONCLUSION

We presented distance mapping approaches to transform 3D head shape information on 2D intensity-encoded images 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), enables the usage of pre-trained CNNs, and preserves patient anonymity. We systematically reduced resolution of the images and showed that using the 2D image structure, low-resolution images can 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 our mapping encoded 3D photogrammetric 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, we publish our Python source code as free and open source software and enable other groups to use our code.

REFERENCES

[1] S. L. Boulet et al., "A population-based study of craniosynostosis in metropolitan Atlanta, 1989," *Amer. J. Med. Genet. Part A*, vol. 146A, no. 8, pp. 984–991, Apr. 2008.

[2] L. French et al., "A population-based study of craniosynostosis," *J. Clin. Epidemiol.*, vol. 43, no. 1, pp. 69–73, Jan. 1990.

[3] J. A. Persing et al., "Virchow and the pathogenesis of craniosynostosis: A translation of his original work," *Plast. Reconstructive Surg.*, vol. 83, no. 4, pp. 738–742, Apr. 1989.

[4] D. Renier et al., "Intracranial pressure in craniostenosis," *J. Neurosurgery*, vol. 57, no. 3, pp. 370–377, Sep. 1982.

[5] K. A. Kapp-Simon et al., "Neurodevelopment of children with single suture craniosynostosis: A review," *Child's Nervous Syst.*, vol. 23, no. 3, pp. 269–281, Jan. 2007.

[6] N. Bannink et al., "Obstructive sleep apnea in children with syndromic craniosynostosis: Long-term respiratory outcome of midface advancement," *Int. J. Oral Maxillofac. Surg.*, vol. 39, no. 2, pp. 115–121, Feb. 2010.

[7] J. A. Fearon et al., "Single sutural craniosynostoses: Surgical outcomes and long-term growth:," *Plast. Reconstructive Surg.*, vol. 123, no. 2, pp. 635–642, Feb. 2009.

[8] S. Nagaraja et al., "Craniosynostosis," *Clin. Radiol.*, vol. 68, no. 3, pp. 284–292, Mar. 2013.

[9] M. Engel et al., "Value of preoperative imaging in the diagnostics of isolated metopic suture synostosis: A riskbenefit analysis," *J. Plastic, Reconstructive Aesthetic Surg.*, vol. 65, no. 9, pp. 1246–1251, Sep. 2012.

[10] K. A. Eley et al., "Black bone" MRI: A potential alternative to CT with three-dimensional reconstruction of the craniofacial skeleton in the diagnosis of craniosynostosis," *Eur. Radiol.*, vol. 24, no. 10, pp. 2417–2426, Oct. 2014.

[11] G. Cacciaguerra et al., "The evolution of the role of imaging in the diagnosis of craniosynostosis: A narrative review," *Children*, vol. 8, no. 9, Aug. 2021, Art. no. 727.

[12] C. Mertens 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," *J. Cranio- Maxillofac. Surg.*, vol. 45, no. 12, pp. 2010–2016, Dec. 2017.

[13] S. Ruiz-Correa et al., "Classifying craniosynostosis deformations by skull shape imaging," in *Proc. IEEE 18th Symp. Comput.-Based Med. Syst.*, 2005, pp. 335–340.

[14] C. S. Mendoza et al., "Personalized assessment of craniosynostosis via statistical shape modeling," *Med. Image Anal.*, vol. 18, no. 4, pp. 635–646, May 2014.

[15] M. Schaufelberger et al., "A radiation-free classification pipeline for craniosynostosis using statistical shape modeling," *Diagnostics*, vol. 12, no. 7, Jun. 2022, Art. no. 1516.

[16] G. de Jong et al., "Combining deep learning with 3D stereophotogrammetry for craniosynostosis diagnosis," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 15346.

[17] S. A. H. Tabatabaei et al., "Automatic detection and monitoring of abnormal skull shape in children with deformational plagiocephaly using deep learning," *Sci. Rep.*, vol. 11, no. 1, Sep. 2021, Art. no. 17970.

[18] S. Agarwal et al., "Classification of Craniosynostosis Images by Vigilant Feature Extraction," in *Advances in Computer Vision and Computational Biology*. Cham, Switzerland: Springer, 2021, pp. 293–306.

[19] S. Lanche et al., "A statistical model of head asymmetry in infants with deformational plagiocephaly," in *Image Analysis*, vol. 4522. Berlin, Germany: Springer, 2007, pp. 898–907.

[20] M. Schaufelberger et al., *A Statistical Shape Model of Craniosynostosis Patients and 100 Model Instances of Each Pathology*. Geneve, Switzerland: Zenodo, 2021.

[21] J. L. Hintze and R. D. Nelson, "Violin plots: A box plot-density trace synergism," *Amer. Statistician*, vol. 52, no. 2, May 1998, Art. no. 181.

[22] M. Waskom, "Seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, Apr. 2021, Art. no. 3021.

[23] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[24] P. Cignoni et al., "MeshLab: An open-source mesh processing tool," in *Proc. Eurograph. Italian Chapter Conf.*, 2008, pp. 129–136.

[25] N. Pietroni, M. Tarini, and P. Cignoni, "Almost isometric mesh parameterization through abstract domains," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 4, pp. 621–635, Jul./Aug. 2010.

[26] G. de Jong et al., "The computed cranial focal point," *J. Cranio- Maxillofac. Surg.*, vol. 43, no. 9, pp. 1737–1742, Nov. 2015.

[27] W. Schroeder et al., *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, USA: Kitware, Inc, 2006.

[28] J. d'Errico, "Matlab central file exchange: Inpaint_nans," Accessed on: Jul. 15, 2022. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/4551-inpaint_nans

[29] Maisevector, *KIT-IBT/CD-Map: Initial Release of Craniosynostosis Distance Maps (CD-Map)*. Geneve, Switzerland: Zenodo, 2022.

[30] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[31] A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[32] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[33] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," Apr. 2017, *arXiv:1704.04861.*

[34] M. Sundararajan et al., "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[35] N. Kokhlikyan et al., "Captum: A unified and generic model interpretability library for PyTorch," Sep. 2020, *arXiv:2009.07896.*

[36] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[38] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[39] P.-J. Kindermans et al., "The (Un)reliability of Saliency Methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Cham, Switzerland: Springer, 2019, pp. 267–280.