Machine Learning in Dentistry

Ching-Chang Ko
Dinggang Shen
Li Wang
Editors

Springer

Machine Learning in Dentistry

Ching-Chang Ko • Dinggang Shen •
Li Wang
Editors

Machine Learning in Dentistry

Springer

Editors
Ching-Chang Ko
Ohio State University
Columbus, OH, USA

Dinggang Shen
School of Biomedical Engineering
ShanghaiTech University
Shanghai, Shanghai, China

Li Wang
University of North Carolina
at Chapel Hill
Chapel Hill, NC, USA

ISBN 978-3-030-71880-0 ISBN 978-3-030-71881-7 (eBook)
https://doi.org/10.1007/978-3-030-71881-7

Preface

Machine Learning in Dentistry

Machine learning (ML) is a branch of artificial intelligence (AI), which is growing fast in both the research community and the general public. In contemporary dentistry, digital technologies such as cone beam computer tomography (CBCT), intra-oral 3D scan, 3D printing, and personalized treatment planning have come to play a larger role in both research and practice. These technologies hold promise for more predictable, objective, and effective treatment while reducing iatrogenic complications. But for change to occur, data scientists and dental experts from various coherent domains must work together to develop big data analytics that have translational values and can be deployed in oral healthcare and life sciences.

In the past, the diagnostic information was collected by clinical interviews, plaster models, and chair-side observations. The information was then interpreted by experts to derive a treatment plan for implementation. The variations in each step of diagnosis, treatment planning, and treatment implementation are huge and purely empirical, depending upon doctor's experience. As such, conventional dental education is largely dependent upon repetitive training of clinician's eyes, hands, and judgment (critical thinking) to minimize the variation in standard care. The amount of time spent in the professional school and in lifelong continued education is excessive, and there are no guarantees that every individual will reach the same standard credential.

Today, the new technologies can replace human's eyes and hands, and AI can teach human learning processes to the machine to close the educational gaps. Many healthcare providers have integrated digital technologies into practice workflows, which reduce the dependence on hand skills and visual recognitions. Computer software is in its infancy to integrate these technologies to improve objective perception, cognition, and action with experience. Nevertheless, ML in dentistry is about developing a personalized precision practice and improving diagnosis and treatment planning from big data, literature, experience, and outcome-dependent learning. These also include the interrogation of large, complex 'omics datasets such as genetic loci and biomarkers for craniofacial disorders.

The purpose of this book is to review the current clinical systems and dental research involving machine learning tools and its associations among various dental specialties. The examples shown in this book represent a sample of the opportunities and challenges for what is possible with the ML approach in

contemporary dentistry. We have invited experts to contribute their review in the following four areas.

Machine Learning for Dental Imaging

Dentists often read radiographs to detect craniofacial anomalies for identifying potential problems. However, mistakes can be made through imaging examination and the diagnostic accuracy is dependent on the experience of the dentist. With a 3D image, this process turns out to be a stressful task because it is based on big data and user experience. By using machine learning to augment the image, diagnosis can be more accurate and objective, and treatments can be personalized.

A computer does not get tired from grueling tasks, and it can take in a great deal of data and process the information very quickly. First, the computer takes in a set of different radiographs labeled for healthy and unhealthy anatomies. Then, once the computer is given data of experts' diagnoses, it can classify the patterns that are associated with certain disease or healthy anatomies. Thus, given a new patient's radiograph, the machine can easily match it with patterns that were found in the training set of expert diagnostics. There are already machine learning programs that are able to detect 2D panoramic features on the market, like denti.ai or dentistry.ai. In this book, we delve deeper into the role the computer does for 3D imaging segmentation and landmark detection using machine learning and deep learning. Chapters 1–5 describe the algorithms and provide examples for augmentation of craniofacial skeletal structures, facial surface recognition, and applications in simulation of orthognathic surgery.

Machine Learning for Oral Diagnosis and Treatment Planning

When conducting a medical diagnosis, a dentist would take in the patient's chief complaint, history, conversation, and radiographs. Some of this information can be descriptive, and because of that, the dentist must be very experienced and in top condition to pay close attention to every word and its nuances from a patient's narrative. Because it assembles these words to make a good diagnosis, dentists may easily misdiagnose their patient. In 2011, dentists were found to have a 43% misdiagnosis rate of oral lesions (Kondori et al.). This number is frighteningly high and indicates room for improvement. In medicine, there are various commercial software for auto-annotations of pathology images. Instead, the book focuses on clinical diagnosis and treatment planning using AI.

Natural language algorithm is part of machine/deep learning, which can support searching for key words and identify frequency and patterns of descriptive statements. Chapters 6–9 provide state-of-the-art AI for orthodontic diagnosis and treatment planning. This includes facial recognition, cephalometric analysis, chart analysis, automated problem list and treatment planning, characterization of craniofacial anomalies, and decision of orthodontic

tooth extraction. The AI could save time and streamline any processes with improved accuracy in the future.

Machine Learning and Dental Designs/Outcome Assessment

Emerging intra-oral scanner and 3D printing have improved workflow of orthodontic practice. To date, the use of digital surface model has had a dramatic expansion from diagnosis to treatment planning, for example orthodontic aligner therapy. Automatic tooth segmentation from the dental arch already benefits the design of restorations, tooth setup and alignment, personalized treatment, and patient-centered outcome assessment. Chapter 10 updates current AI technologies for tooth isolation from the 3D surface model. Dental specialties have been in the forefront of using artificial intelligence analytic strategies to leverage the advantages of "big data" to personalize treatment interventions, such as aligner therapy. Chapter 11 provides an overview of machine learning and how it can be applied in assessing outcomes. Specifically, we will focus on recent advances in craniofacial genomics and outcomes pertaining to images.

Machine Learning Supporting Dental Research

Oral systemic health is tightly associated with good dental research. Clinical research is to develop evidence-based guidelines and their potential impact on the health and well-being of large portions of the population. The translational value of these studies relies significantly on the effectiveness of the systemic review. Chapter 12 provides a review of using ML for a systemic review. The authors also discuss about the shortcomings and potential pitfalls of employing artificial intelligence (AI) to a review process that has been mostly driven by human experts in the past.

Human genomics was one of the first areas of research to use big data. In Chap. 13, the researchers introduce how the machine learning and deep learning algorithms are used in dental omics studies for association analysis between SNPs and complex diseases, CNV and SNV calling, and DNA methylation data. In the end, we discuss the potential developing direction of machine (deep) learning in biomechanics in oral health (Chap. 14).

We thank Springer and all the contributors for their collective articles, Dr. Tai-Hsien Wu and Dr. Chunfeng Lian for the liaison task to communicate with the contributors, and Ms. Devan for careful management of the submissions. Particular thanks also go to Drs. Pastewait, Piers, and Chien and Ms. Zheng for reviewing the manuscripts. We believe this book is just the beginning of a series of books on the topic of AI in contemporary dentistry.

Columbus, OH — Ching-Chang Ko
Shanghai, China — Dinggang Shen
Chapel Hill, NC — Li Wang

Contents

Part I

Machine Learning for Dental Imaging

1 Machine Learning for CBCT Segmentation of Craniomaxillofacial Bony Structures

Chunfeng Lian, James J. Xia, Dinggang Shen, and Li Wang

1.1 Introduction

Cone-beam computed tomography (CBCT) has been routinely used in the diagnosis and treatment of patients with craniomaxillofacial (CMF) deformities. It has lower dose and lower cost than the conventional spiral multi-slice computed topography (MSCT) [21]. Accurate CBCT segmentation of CMF bones (i.e., maxilla and mandible) to construct a precise three-dimensional (3D) skeletal model is necessary to quantitatively assess various deformities [32]. However, the segmentation of dentomaxillofacial CBCT images is challenging, mainly considering that (i) CBCT images usually contain considerable artifacts, (ii) CBCT has low signal-to-noise ratio due to the utilization of low dose radiation, and (iii) the maxillary and mandibular teeth are cross-sectionally overlapped in CBCT images due to the requirement of maximum intercuspation for diagnostic purpose. Due to these challenges as well as the large size of CBCT volumes, manually annotating CMF bones is labor intensive and experience dependent, which suggests the importance of developing automated methods to this end.

Conventional automated methods for CBCT segmentation include thresholding and morphological methods, iterative methods, and model-based methods. The thresholding methods [11] usually adopt simple intensity threshold and craniofacial morphological criteria to extract bones, which are sensitive to CBCT image artifacts. The semi-automated iterative methods [15, 25] combine automated segmentation with experts' manual intervention for bone annotation. Their performance is also largely influenced by image artifacts, e.g., those caused by dental implants [28]. The model-based methods [9, 12, 14, 28] typically conduct segmentations based on statistical shape models (SSMs) or atlases. For example, Gollmer et al. [9] proposed to minimize an objective function constrained by statistical shape prior for mandible segmentation. Considering that such kinds of prior only work for objects with relatively simple shapes (e.g., mandible), rather than those with complex shapes (e.g., maxilla), Wang et al. [28] proposed a patch-based sparse-coding method for the robust extraction of both maxilla and

C. Lian · D. Shen · L. Wang (✉)
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: chunfeng_lian@med.unc.edu; dgshen@med.unc.edu; li_wang@med.unc.edu

J. J. Xia
Department of Oral and Maxillofacial Surgery, The Methodist Hospital Research Institute, Houston, TX, USA
e-mail: JXia@houstonmethodist.org

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_1

mandible from soft tissues. However, since the patch-based sparse coding relies on deformable registration (between target and multiple atlases) as well as per-voxel optimization, the model-based method proposed in [28] is time consuming at both the training and inference stages.

In recent years, machine learning algorithms have been successfully applied in diverse medical image computing tasks, such as reconstruction/synthesis [10, 34], registration [5, 24], segmentation [18,19,23,30], detection [8,26,33], and diagnosis [1, 16, 17]. For example, as a nonparametric ensemble learning method, random forest (RF) [4] and its variants are showing competitive performance in many segmentation tasks [7, 13, 30, 31], by learning an ensemble model of decision trees from randomly sampled imaging features and training samples. Inspired by those successful applications, a novel RF-based sequential model was proposed in [29], which combines high-quality (and gradually refined) anatomical prior with appearance imaging features to construct a series of RFs for efficient and fully automated CBCT segmentation of maxillas and mandibles from soft tissues. Briefly, such a prior-guided sequential RF method consists of a training step and a test step. In the training step, both the appearance features from CBCT images and the contextual features from the initial segmentation probability maps (of maxilla and mandible) are extracted, based on which sequential RFs are trained to select discriminative features for the iterative refinement of segmentation probability maps. In the test step, using the same initial inputs extracted from any test image, the learned sequential RFs are orderly applied to refining the probability maps of the maxilla and mandible for the test image.

In this chapter, we will introduce this prior-guided sequential RF method in detail. The remaining part of this chapter is organized as follows. In Sect. 1.2, we present the fundamental background of random forests and imaging feature extraction. In Sect. 1.3, we introduce how the prior-guided sequential RF works. In Sect. 1.4, we present some experimental results on a dataset of CBCT images for patients with dentofacial deformities. The chapter is concluded in Sect. 1.5.

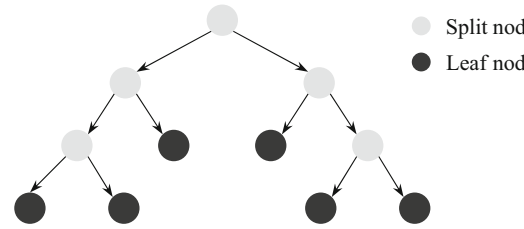


Fig. 1.1 A simplified diagram of binary decision tree with the depth equal to 3

1.2 Background

1.2.1 Random Forest

As a general supervised learning method with good efficiency and scalability, random forest can be basically regarded as an average of multiple binary decision trees. Each of those decision trees consists of two types of nodes, i.e., split nodes and leaf nodes. As shown in Fig. 1.1, the split nodes route the incoming data either to their left or right child node. Each of the leaf nodes is a terminal of serial binary splits, storing the information of training samples routed to it (i.e., the ratios of different categories). All decision trees are trained independently with randomly sampled training data (i.e., bootstrap samples), and each of their split nodes is associated with a specific split function in terms of a randomly sampled feature/descriptor subset. In the remaining part of this subsection, we will introduce some fundamental background on random forests, including the construction of a binary decision tree, and the training of a random forest for classification.

1.2.1.1 Construction of a Binary Decision Tree

Given a set of training samples (e.g., CBCT voxels) and a set of features that can be extracted (e.g., Haar-like features), a decision tree is trained to recursively route the samples to different leaf nodes, by maximizing the purity of each subset during each binary split. In the case of classification (e.g., predicting whether a CBCT voxel is from maxilla, mandible, or background), the purity of a subset is defined based on the labels

of training samples assigned to it. That is, the respective split function should encourage the partition of a training set into subsets with the same label.

Let $\mathcal{S}$ be a set of training samples coming to a split node, and each sample has a class label in $\{1, \ldots, C\}$. A random tuple of features is indexed by a set Φ. The split at the node works as

$$\arg \max_{\phi \in \Phi,\, \mathcal{S}_\mathrm{L},\, \mathcal{S}_\mathrm{R}} \frac{1}{|\mathcal{S}_\mathrm{L}|} \sum_c p_c^\mathrm{L} \log p_c^\mathrm{L} + \frac{1}{|\mathcal{S}_\mathrm{R}|} \sum_c p_c^\mathrm{R} \log p_c^\mathrm{R}, \tag{1.1}$$

where $\mathcal{S}_\mathrm{L} = \{s \in \mathcal{S} | f(s, \phi) \leq t_\phi\}$ and $\mathcal{S}_\mathrm{R} = \{s \in \mathcal{S} | f(s, \phi) > t_\phi\}$ are the binary sample subsets from $\mathcal{S}$, with $s \in \mathcal{S}$ being a specific training sample. The coefficients p_c^L and p_c^R denote the ratios of the c-th class samples in $\mathcal{S}_\mathrm{L}$ and $\mathcal{S}_\mathrm{R}$, respectively. The operation $f(s, \phi)$ denotes the extraction of the ϕ-th feature for the sample s, and t_ϕ is the corresponding threshold that is randomly determined.

From the root node (i.e., the first split node) to the leaf nodes, the above step is recursively operated by randomly sampling a feature subset Φ at each split node, until one of the following criteria is satisfied [7]. That is, (i) the decision tree reaches a predefined maximum tree depth (e.g., the depth of tree in Fig. 1.1 is 3), or (ii) the training subset coming to a node is too small to be further split, or (iii) the purity for a node is above a predefined threshold. When a split stops, the corresponding node is made as a leaf node, and the label information regarding the training set arriving there is stored.

After training, any test sample comes to the root node and is then recursively routed to the subsequent split nodes according to the associated split functions, until it reaches a leaf node. The probability of the test sample belonging to any class is determined by the label information of the training samples stored in the respective leaf node.

1.2.1.2 Construction of a Random Forest

A random forest is an ensemble of binary decision trees. The trees are trained independently by randomly sampling different training subsets with replacement from the original training set (i.e., *bootstrap* sampling) and randomly selecting features, which can largely increase the model diversity to improve the generalization capacity [4, 7].

The leaf node of each trained tree stores the distribution of training samples (in terms of class labels) routed to it. During inference, each test sample will arrive at a leaf node in each tree, and the respective label distribution of training samples is used to infer the class probability of this test sample. We finally normalize the label distribution values from all trees as the class likelihood for each test sample, and the label is determined as the class with the maximum likelihood.

It is worth mentioning that, compared with other classification models (e.g., support vector machine), random forests have the advantage of high efficiency for processing data with high-dimensional input features during inference. This is mainly because each test sample only goes through one path from the root node to a specific leaf node in each tree, which means we only need to extract the features that are used in this specific path (instead of all input features used during training) for this test sample.

1.2.2 Feature Extraction

In the communities of computer vision and medical image computing, diverse kinds of features have been engineered according to the properties of different tasks. The most commonly used features include the scale-invariant feature transform (SIFT) [22], Haar-like features [20], local binary pattern (LBP) [2], and histogram of oriented gradients (HOG) [6].

Considering the computational efficiency, we use 3D Haar-like features [29] to construct the prior-guided sequential random forests. For each voxel in a volume, its Haar-like features are calculated in terms of a 3D patch or two asymmetric 3D patches that are randomly displaced from the voxel. Mathematically, given a voxel v and its two randomly displaced patches P_1 and P_2, the Haar-

like features are defined as

$$f(v) = \frac{1}{|P_1|}\sum_{u\in P_1} I(u) - b\frac{1}{|P_2|}\sum_{w\in P_2} I(w), \quad (1.2)$$

where $I(u)$ denotes the image intensity of a voxel u and $b \in \{0, 1\}$. In [29], the Haar-like features were extracted from both the original CBCT image and segmentation probability maps, and the patches P_1 and P_2 were sampled in a $17 \times 17 \times 17$ neighborhood. As will be introduced in the next section, the same kind of features from different inputs actually provide complementary but different information.

1.3 Prior-Guided Sequential Random Forests

In this chapter, we regard the CBCT segmentation of the maxilla and mandible as a voxel-wise classification task. That is, imaging features are extracted for each voxel to train a classification model, i.e., the prior-guided sequential RFs, which predicts the probabilities of the respective voxel belonging to the maxilla, mandible, and background. The final segmentation is produced by assigning the label with the largest probability value at each voxel location.

The same as other supervised machine learning approaches, the prior-guided sequential RF method consists of a training step and a test step, with the details described below.

1.3.1 Training

Given a set of atlases (i.e., expert-segmented CBCT volumes) and a set of training samples, the segmentation model is constructed following the pipeline shown in Fig. 1.2. The training step contains three important components: (1) estimation of the initial probability maps for each training image, (2) extraction of both contextual and appearance features from the probability maps and original image, respectively, and (3) construction of a series of auto-context RFs to gradually refine the probability maps for the training image.

1.3.1.1 Initial Probability Maps

Multiple expert-segmented CBCT volumes are used as the atlases to estimate the initial probability maps of the maxilla and mandible. Those initial probability maps can provide important anatomical priors to guide the construction of an automated segmentation model [36]. Specifically, the atlases (along with the corresponding manual segmentation maps) are rigidly aligned onto each training image. After that, for each voxel in a training image, we count the numbers of labels (i.e., maxilla, mandible, and background) at the corresponding locations from all available atlases and then normalize them (followed by Gaussian smooth with $\delta = 2$ mm [35]) as the initial probabilities that describe the voxel belonging to different categories. This simple and efficient strategy can produce reasonable initial probability maps, such as those shown in Fig. 1.2.

1.3.1.2 Complementary Imaging Features

To comprehensively describe each CBCT image voxel for the construction of a reliable voxel-wise classification model, we extract the imaging features (i.e., the random Haar-like features [27]) from both the original CBCT image and the respective segmentation probability maps estimated by the preceding step. It is worth noting that imaging features from the original image and probability maps provide different and complementary information. Generally, the original image provides low-level appearance information, while the segmentation probability maps provide high-level contextual/semantic information. The classifier constructed with the features extracted from both the original image and the probability maps is typically called the auto-context model, which has shown very successful applications in both the communities of computer vision [3] and medical image computing [19, 30].

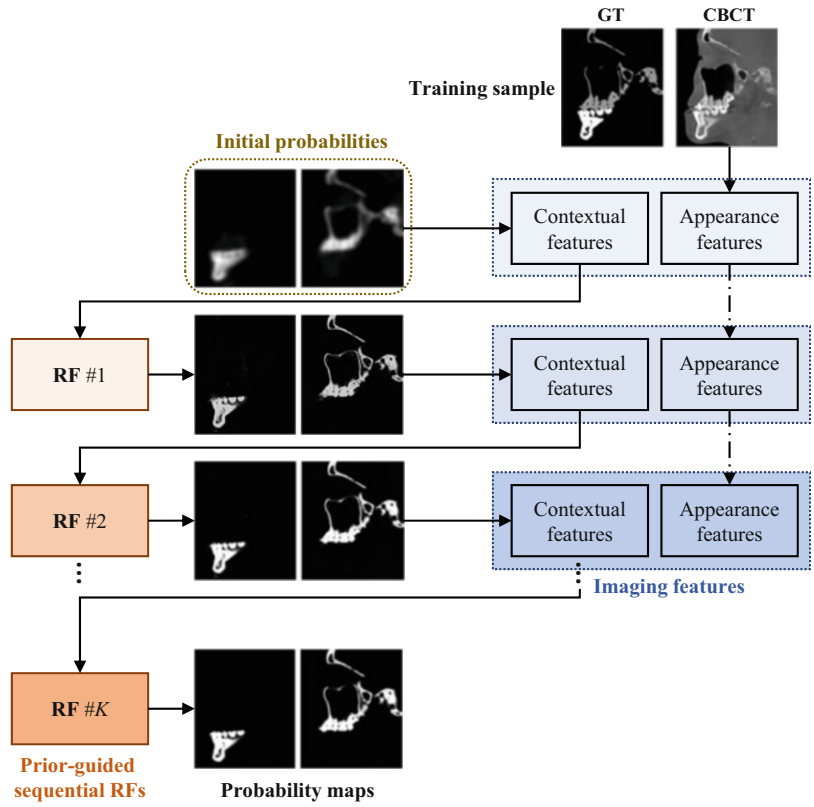


Fig. 1.2 The schematic diagram of the training of prior-guided sequential RFs for CMF bony segmentation in CBCT images. There are three important components, i.e., the definition of initial probability maps for a training sample, the extraction of imaging features, and the construction of a series of auto-context RFs. The GT stands for ground truth. (Please note this figure was copied from [29])

1.3.1.3 Sequential RFs

The contextual and appearance imaging features for each voxel are combined as a single feature vector for the training of an auto-context classification model. Instead of other classic classifiers (e.g., support vector machine and multi-layer perceptron), random forest is used here primarily due to the reason that it can effectively handle large-scale training samples (i.e., image voxels) and can efficiently select the discriminative features from high-dimensional input feature space.

To gradually improve the quality of contextual features for more accurate segmentation, multiple auto-context RFs are trained in a serial manner until convergence. That is, the contextual features are repeatedly updated by extracting them from the segmentation probability maps predicted by the latest auto-context RF, which are further combined with the appearance features for the training of a new auto-context RF. Generally, the segmentation performance of the whole model can converge within limited iterations.

1.3.2 Inference

After training, the prior-guided sequential RFs are orderly applied on test CBCT images to produce probability maps of maxillas and mandibles with increasingly improved qualities, such as the example shown in Fig. 1.3. Specifically, for an input CBCT image shown in Fig. 1.3a, the same atlases used in the training stage are rigidly aligned onto that for the generation of the initial probability maps in Fig. 1.3b. Then, contextual and appearance features are extracted from the probability maps and original CBCT image, respectively, which are fed into the first auto-context RF to produce the refined probability maps in Fig. 1.3c. The updated contextual features are further extracted from Fig. 1.3c and combined with appearance feature to train the subsequent auto-context RF, generating the predictions in Fig. 1.3d. This procedure is repeated until the output of the final probability maps in Fig. 1.3e. Here, the exam-

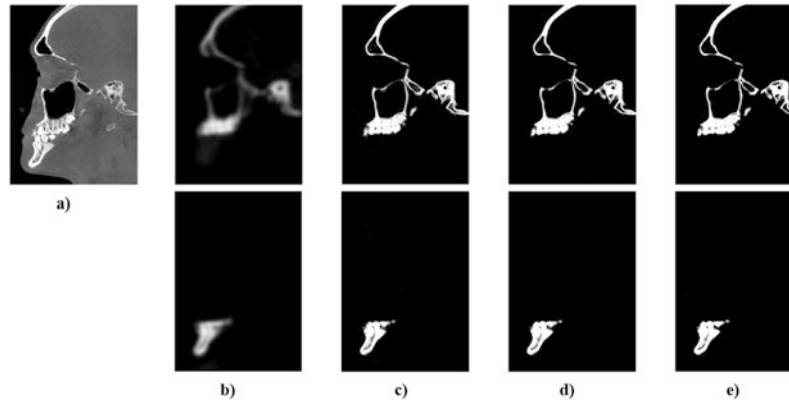


Fig. 1.3 The gradual refinement of the probability maps of maxilla and mandible for an input CBCT image using the prior-guided sequential RFs. (Please note this figure was copied from [29]). (**a**) Input CBCT (**b**) Initial probability (**c**) RF #1 prediction (**d**) RF #2 prediction (**e**) RF #3 predictions

ple prior-guided sequential model contains three auto-context RFs.

1.4 Experiments

1.4.1 Dataset and Pre-processing

The studied dataset consists of 30 CBCT scans for different patients with nonsyndromic dentofacial deformities, treated with a double-jaw orthognathic surgery. These patients include 12 males and 18 females, and their ages were 24 ± 10 years. Their CBCT scans were acquired by an i-CAT machine with a matrix of 400×400, and exposure time of 40 s, resulting in 3D volumes with isotropic resolution (voxel size: $0.4\,mm \times 0.4\,mm \times 0.4\,mm$). All of the CBCT images were HIPAA de-identified, and the ground-truth bony segmentations were manually annotated by one experienced CMF surgeon using Mimics 10.01 software (Materialise NV, Leuven, Belgium). Manually labeling one volume required about 12 h.

During training, 5, 000 voxel locations were sampled from each training image. Using each voxel location as the center, 10, 000-D Haar-like features were extracted from $17 \times 17 \times 17$ local patches in the original image and segmentation probability maps, respectively. After training, each voxel location in a test image was orderly used as the center to extract imaging features, which were fed into the trained model to predict the probabilities of the maxilla and mandible for the respective voxel. The automated segmentation of one subject required about 20 min, which was significantly faster than manual annotation.

1.4.2 Experimental Setting

The evaluation of the prior-guided sequential random forest method was performed via leave-one-out cross-validation. That is, by iterating 30 times, each of the 30 subjects was orderly selected as a test sample and the remaining subjects were used as the training set. The tuning parameters of a random forest (RF) were determined on the training set via nested cross-validation. Specifically, each RF was constructed by 40 deep classification trees, and the depth of each tree was not larger than 100.

Using the manual annotations as the reference, the automated segmentation performance was quantified by multiple metrics, including (1) dice ratio (DR) that is proportional to the overlaps between the manual and automated segmentations, (2) average surface distance (ASD), and (3) Hausdorff distance (HD) that is inversely proportional to the consistency between the surfaces of the manual and automated segmentations. More precisely, given segmentations A and B, the dice ratio is defined as

$$\mathrm{DR}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{1.3}$$

where $|A|$ denotes the number of voxels in A and $A \cap B$ denotes the overlap between A and B. The

average surface distance is defined as

$$\mathrm{SDE}(A, B) = \frac{1}{2}\left(\frac{\sum_{a\in \mathrm{surf}(A)} \mathrm{dist}(a, B)}{|\mathrm{sruf}(A)|} + \frac{\sum_{b\in \mathrm{surf}(B)} \mathrm{dist}(b, A)}{|\mathrm{sruf}(B)|}\right), \tag{1.4}$$

where $\mathrm{surf}(A)$ represents the set of points on the surface of A and $\mathrm{dist}(a, B)$ denotes the shortest distance from point a to the surface of B. The Hausdorff distance is defined as

$$\mathrm{HD}(A, B) = \max\left(\max_{a\in A} \mathrm{dist}(a, B), \max_{b\in B} \mathrm{dist}(b, A)\right). \tag{1.5}$$

The prior-guided sequential random forest method was compared with other state-of-the-art methods to demonstrate its superior performance and the effectiveness of its important components. The competing methods include (1) majority voting (MV), (2) patch-based sparse representation (SR) [28], and (3) sequential random forests without prior. Notably, the MV method was also used to generate the initial probabilities for the prior-guided sequential random forest method. The sequential random forest did not include the initial probability maps for the construction of the RF in the first iteration.

1.4.3 Segmentation Results of Maxilla and Mandible

A typical subject and the corresponding segmentations produced by different automated methods, as well as a manual segmentation by an experienced CMF surgeon, are shown in Fig. 1.4, where the first row presents the overall segmentations and the second row presents the zoomed-in visualization of local segmentation details. From Fig. 1.4, we can have the following observations. *First*, the result by MV is not as accurate as other more advanced methods (e.g., SR and the proposed method) due to possible errors during affine registration. However, it is worth noting that the segmentation produced by MV is reasonable, which is sufficient to be used as the guidance for the prior-guided sequential random forests. *Second*, SR produced better results than MV, due to the utilization of a more accurate nonlinear registration method for the alignment between atlases and target subject as well as the optimization of sparse learning parameters. However, the zoomed-in visualization shows that SR cannot properly separate the maxilla and mandible at the closed-bite position. *Third*, the sequential RF without prior resulted in considerable isolated false predictions, which suggests the importance of the anatomical prior provided by the initial probability maps for the proposed method. By including the prior, the proposed method yielded the best automated segmentation.

Besides the qualitative evaluation, the automated segmentation performance of different methods was also quantitatively compared in terms of three metrics (i.e., DR, ASD, and HD), with the average results (mean $\pm$ standard deviation) of all subjects summarized in Table 1.1. It can be seen that the quantitative evaluation in Table 1.1 is consistent with the qualitative evaluation in Fig. 1.4. In other words, the prior-guided sequential random forest method led to the best segmentation results in terms of all metrics.

1.4.4 Segmentation Results of Teeth

In addition to the segmentation of the whole CMF model, the performance of the prior-guided sequential random forest was also analyzed in the separation of maxillary (i.e., upper) and mandibular (i.e., lower) teeth. In Fig. 1.5, the teeth (including incisors, canines, and molars) segmentation results by SR, prior-guided sequential RFs, and a CMF surgeon (i.e., the ground truth) are compared. The included examples have different appearances, varying from open mouth to closed bite. It can be seen that the prior-guided sequential random forest method led to more consistent results with the ground truth across all cases, compared with the SR method. Notably, the improvement brought by the prior-guided sequential random forest is even larger for the challenging closed-bite cases, which further demonstrates its effectiveness.

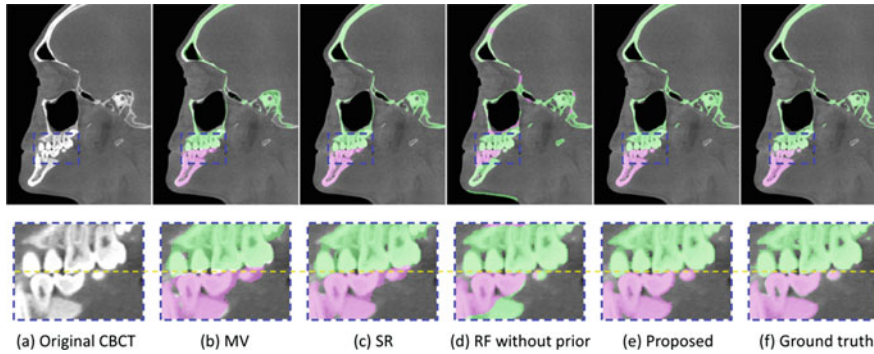


Fig. 1.4 The segmentation results of (**a**) a typical example produced by (**b**) MV, (**c**) SR, (**d**) sequential RFs without prior, (**e**) the proposed prior-guided sequential RFs, and (**f**) the ground-truth manual annotations. (Please note this figure was copied from [29])

Table 1.1 Segmentation results (mean ± standard deviation) quantified under leave-one-out cross-validation for four different automated methods. The performance was jointly evaluated by DR (dice ratio), ASD (average surface distance, in *mm*), and HD (Hausdorff distance, in *mm*). The best results are bolded. (Please note this table was copied from [29])

		MV	SR	RF	Prior-guided RF
DR	*Mandible*	0.83 ± 0.04	0.92 ± 0.03	0.51 ± 0.12	$\mathbf{0.94 \pm 0.02}$
	Maxilla	0.75 ± 0.04	0.88 ± 0.02	0.39 ± 0.14	$\mathbf{0.91 \pm 0.03}$
ASD	*Mandible*	1.21 ± 0.25	0.62 ± 0.02	3.42 ± 1.21	$\mathbf{0.42 \pm 0.15}$
HD	*Mandible*	3.65 ± 1.53	0.95 ± 0.02	4.74 ± 2.56	$\mathbf{0.74 \pm 0.25}$

The segmentation performance of teeth was also quantitatively evaluated, with the results presented in Table 1.2. We can observe that the prior-guided sequential RF method outperformed the SR method by a large margin in the annotation of upper and lower teeth.

1.4.5 Importance of Prior and Sequential Learning

The utilization of initial probability maps to provide anatomical prior and the learning of sequential random forests for probability map refinement are two critical components of the method introduced in this chapter. To more clearly demonstrate the effectiveness of these two components, the CMF models generated by automated segmentations for a typical example are compared with the ground truth in Fig. 1.6. Specifically, Fig. 1.6a presents the CMF surface generated by the sequential random forests (at the third iteration) without prior, Fig. 1.6b presents the surface corresponding to the initial probabilities for the sequential random forests with prior, Fig. 1.6c presents the surface generated by the prior-guided sequential random forests at the third iteration, and Fig. 1.6d presents the ground-truth surface manually defined by the surgeon. One can see that the results presented in Fig. 1.6a are poor, as considerable false predictions exist in both the maxillary and mandibular regions. These false predictions are mainly due to the fact that the two regions have a very similar appearance in the CBCT images, which highlights the importance of the anatomical prior in helping to distinguish between them. On the other hand, we can also observe that, compared with Fig. 1.6b, the local details of Fig. 1.6c are more consistent with the ground truth, which verifies the effectiveness of sequential learning in refining the output segmentations.

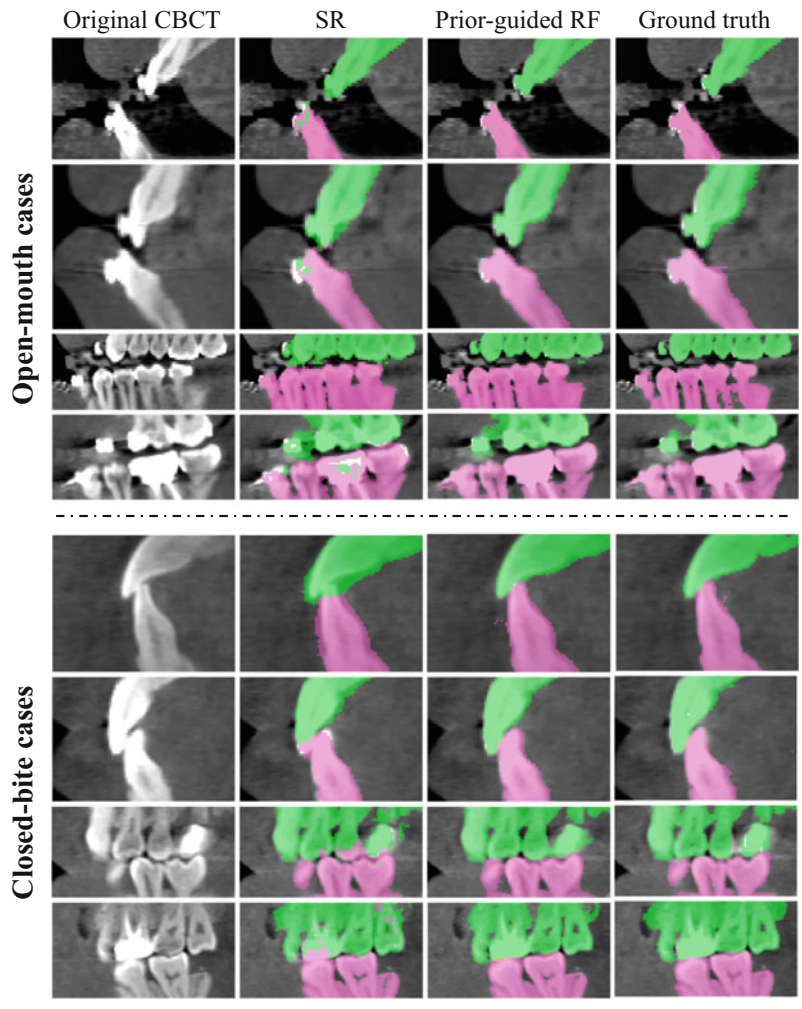


Fig. 1.5 Typical segmentation results of teeth with different appearance (i.e., open mouth or closed bite) by SR, prior-guided sequential RF, and experienced CMF surgeon (i.e., ground truth). (Please note this figure was copied from [29])

Table 1.2 Segmentation results (mean ± standard deviation) of upper and lower teeth quantified under leave-one-out cross-validation for four different automated methods. The performance was jointly evaluated by DR (dice ratio), ASD (average surface distance, in *mm*), and HD (Hausdorff distance, in *mm*). The best results are bolded. (Please note this table was copied from [29])

		SR	Prior-guided RF
DR	*Upper*	0.89 ± 0.03	$\mathbf{0.93 \pm 0.02}$
	Lower	0.92 ± 0.02	$\mathbf{0.96 \pm 0.02}$
ASD	*Upper*	0.74 ± 0.41	$\mathbf{0.35 \pm 0.15}$
	Lower	0.72 ± 0.34	$\mathbf{0.31 \pm 0.10}$
HD	*Upper*	1.36 ± 0.35	$\mathbf{0.67 \pm 0.21}$
	Lower	1.27 ± 0.32	$\mathbf{0.62 \pm 0.19}$

1.5 Conclusion

A prior-guided sequential random forest method has been introduced in this chapter, which was proposed in [29] for fully automated segmentation of CMF bony structures from CBCT images. This method adopts initial probability maps (of maxillas and mandibles), estimated by atlas-based majority voting, as the prior to provide contextual guidance for the training of the subsequent classification model. The Haar-like features are extracted from both the original CBCT

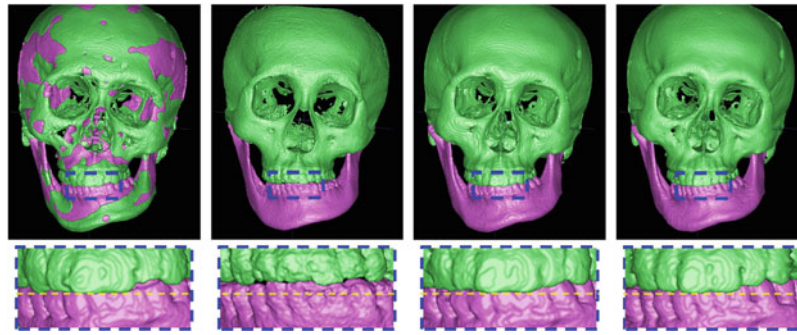


Fig. 1.6 The CMF surfaces generated by the sequential random forest, where (**a**) to (**d**) present the results for the sequential RFs without prior at the 3rd iteration, the initial probability maps for the prior-guided sequential RFs, the prior-guided sequential RFs at the 3rd iteration, and the ground truth. (Please note this figure was copied from [29])

image and the probability maps, which provide appearance and contextual information, respectively, to construct a random forest for the classification of each voxel. A series of random forests were trained sequentially, where each random forest extracts the contextual features from the probability maps predicted by the preceding random forest, resulting in gradually refined segmentation probabilities. The prior-guided sequential random forest has been validated on 30 CBCT volumes for patients with nonsyndromic dentofacial deformities. The experimental results have demonstrated the superior performance of this method compared with previous methods. The effectiveness of its important components has also been verified.

Acknowledgments The authors of this chapter were supported in part by NIH/NIDCR Research Grants (R01 DE027251 and R01 DE022676).

References

1. Adeli E, Shi F, An L, Wee CY, Wu G, Wang T, Shen D. Joint feature-sample selection and robust diagnosis of Parkinson's disease from MRI data. NeuroImage. 2016;141:206–19.
2. Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns. In: European Conference on Computer Vision (ECCV). Springer; 2004. p. 469–81.
3. Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell. 2002;24(4):509–22.
4. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
5. Cao X, Yang J, Gao Y, Wang Q, Shen D. Region-adaptive deformable registration of CT/MRI pelvic images via learning-based image synthesis. IEEE Trans Image Process. 2018;27(7):3500–12.
6. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). 2005.
7. Gao Y. Accurate segmentation of CT pelvic organs via incremental cascade learning and regression-based deformable models. Ph.D. thesis, The University of North Carolina at Chapel Hill. 2016.
8. Gao Y, Shen D. Collaborative regression-based anatomical landmark detection. Phys Med Biol. 2015;60(24):9377.
9. Gollmer ST, Buzug TM. Fully automatic shape constrained mandible segmentation from cone-beam CT data. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2012. p. 1272–5.
10. Gramfort A, Poupon C, Descoteaux M. Denoising and fast diffusion imaging with physically constrained sparse dictionary learning. Med Image Anal. 2014;18(1):36–49.
11. Hassan BA. Applications of cone beam computed tomography in orthodontics and endodontics. 2010.
12. Kainmueller D, Lamecker H, Seim H, Zinser M, Zachow S. Automatic extraction of mandibular nerve and bone from cone-beam CT data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; 2009. p. 76–83.

13. Kamiya N, Li J, Kume M, Fujita H, Shen D, Zheng G. Fully automatic segmentation of paraspinal muscles from 3D torso CT images via multi-scale iterative random forest classifications. Int J Comput Assist Radiol Surg. 2018;13(11):1697–706.
14. Lamecker H, Kainmueller D, Zachow S, et al. Automatic detection and classification of teeth in CT data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; 2012. p. 609–16.
15. Le BH, Deng Z, Xia J, Chang YB, Zhou X. An interactive geometric technique for upper and lower teeth segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; 2009. p. 968–75.
16. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE Trans Pattern Anal Mach Intell. 2018;42:880–93.
17. Lian C, Ruan S, Denœux T, Jardin F, Vera P. Selecting radiomic features from FDG-PET images for cancer treatment outcome prediction. Med Image Anal. 2016;32:257–68.
18. Lian C, Ruan S, Denœux T, Li H, Vera P. Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions. IEEE Trans Image Proces. 2018;28(2):755–66.
19. Lian C, Zhang J, Liu M, Zong X, Hung SC, Lin W, Shen D. Multi-channel multi-scale fully convolutional network for 3D perivascular spaces segmentation in 7T MR images. Med Image Anal. 2018;46:106–17.
20. Lienhart R, Maydt J. An extended set of Haar-like features for rapid object detection. In: IEEE International Conference on Image Processing, vol. 1. IEEE; 2002, p. I–I.
21. Loubele M, Bogaerts R, Van Dijck E, Pauwels R, Vanheusden S, Suetens P, Marchal G, Sanderink G, Jacobs R. Comparison between effective radiation dose of CBCT and MSCT scanners for dentomaxillofacial applications. Eur J Radiol. 2009;71(3):461–8.
22. Lowe DG, et al. Object recognition from local scale-invariant features. In: the Seventh IEEE International Conference on Computer Vision (ICCV), vol. 99; 1999. p. 1150–7.
23. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; 2015. p. 234–41.
24. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: A survey. IEEE Trans Med Imaging. 2013;32(7):1153.
25. Suebnukarn S, Haddawy P, Dailey M, Cao D. Interactive segmentation and three-dimension reconstruction for cone-beam computed-tomography images. NECTEC Techn J. 2008;8(20):154–61.
26. Torosdagli N, Liberton DK, Verma P, Sincan M, Lee JS, Bagci U. Deep geodesic learning for segmentation and anatomical landmarking. IEEE Trans Med Imaging. 2018;38(4):919–31.
27. Viola P, Jones MJ. Robust real-time face detection. Int J Comput Vis. 2004;57(2):137–54.
28. Wang L, Chen KC, Gao Y, Shi F, Liao S, Li G, Shen SG, Yan J, Lee PK, Chow B, et al. Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization. Med Phys. 2014;41(4):043503.
29. Wang L, Gao Y, Shi F, Li G, Chen KC, Tang Z, Xia JJ, Shen D. Automated segmentation of dental CBCT image with prior-guided sequential random forests. Med Phys. 2016;43(1):336–46.
30. Wang L, Gao Y, Shi F, Li G, Gilmore JH, Lin W, Shen D. LINKS: Learning-based multi-source integration framework for segmentation of infant brain images. NeuroImage. 2015;108:160–72.
31. Wang Z, Wei L, Wang L, Gao Y, Chen W, Shen D. Hierarchical vertex regression-based segmentation of head and neck CT images for radiotherapy planning. IEEE Trans Image Process. 2017;27(2):923–37.
32. Yuan P, Mai H, Li J, Ho DCY, Lai Y, Liu S, Kim D, Xiong Z, Alfi DM, Teichgraeber JF, et al. Design, development and clinical validation of computer-aided surgical simulation system for streamlined orthognathic surgical planning. Int J Comput Assist Radiol Surg. 2017;12(12):2129–43.
33. Zhang J, Gao Y, Wang L, Tang Z, Xia JJ, Shen D. Automatic craniomaxillofacial landmark digitization via segmentation-guided partially-joint regression forest model and multiscale statistical features. IEEE Trans Biomed Eng. 2015;63(9):1820–9.
34. Zhang Y, Yap PT, Chen G, Lin W, Wang L, Shen D. Super-resolution reconstruction of neonatal brain magnetic resonance images via residual structured sparse representation. Med Image Anal. 2019;55:76–87.
35. Zikic D, Glocker B, Criminisi A. Atlas encoding by randomized forests for efficient label propagation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer; 2013. p. 66–73.
36. Zikic D, Glocker B, Criminisi A. Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. Med Image Anal. 2014;18(8):1262–73.

2 Machine Learning for Craniomaxillofacial Landmark Digitization of 3D Imaging

Jun Zhang, Mingxia Liu, Li Wang, Chunfeng Lian, and Dinggang Shen

2.1 Background

Craniomaxillofacial (CMF) deformities involve congenital and acquired deformities of the head and face (as shown in Fig. 2.1). Because of these malformations, daily life can be seriously affected. Specifically, jaw deformity is the most common type of CMF deformity, and orthognathic surgery is the procedure to correct it.

In diagnosis and treatment planning, computed tomography (CT) scans, including multi-slice CT (MSCT) and cone-beam CT (CBCT), are typically processed using the following procedures: segmentation, reconstruction of 3D models, placing anatomical landmarks on the 3D models (called *landmark digitization*), and quantitative measurements based on the landmarks (called *cephalometry*). Compared with MSCT, CBCT has less scanning time, reduced radiation dose, lower cost, and the potential for in-office imaging, which is increasingly being used for CMF patients. With regard to CT scan processing, accurate landmark digitization is a critical step to quantitatively assess the CMF anatomy. In current clinical practice, almost all anatomical CMF landmarks are manually digitized on 3D models, which is time-consuming. The goal is to have an automatic landmark digitization method, which can incorporate machine learning techniques to achieve rapid and accurate landmark identification. Moreover, landmark digitization has several clinical applications (as shown in Fig. 2.2). It can help with cephalometric analysis as well as teeth segmentation [1]. In addition, landmark digitization can be used for surgical planning by generating appropriate reference models [2]. It should also be noted that landmark-based image analysis has been attracting increasing attention for various medical applications, including lesion detection [3], region selection [4], image registration [5], disease diagnosis [6–9], and disease prognosis [10, 11].

To date, there is no effective method that allows automatic landmark digitization in clinical

J. Zhang (✉)
Tencent AI Lab, Shenzhen, Guangdong, China

M. Liu · L. Wang · D. Shen
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

C. Lian
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shannxi, China

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_2

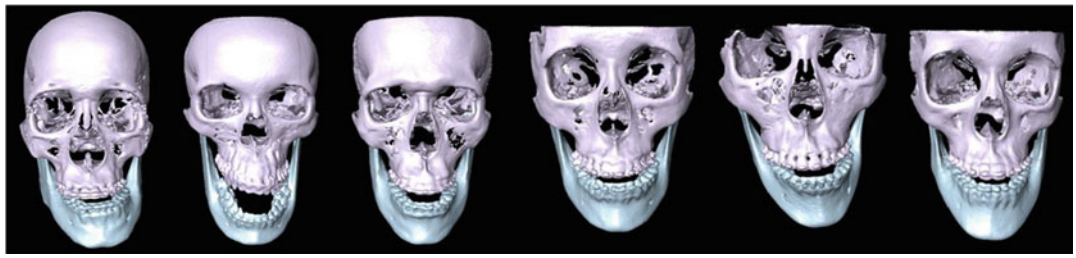


Fig. 2.1 Patients suffer severe deformities of the head and face

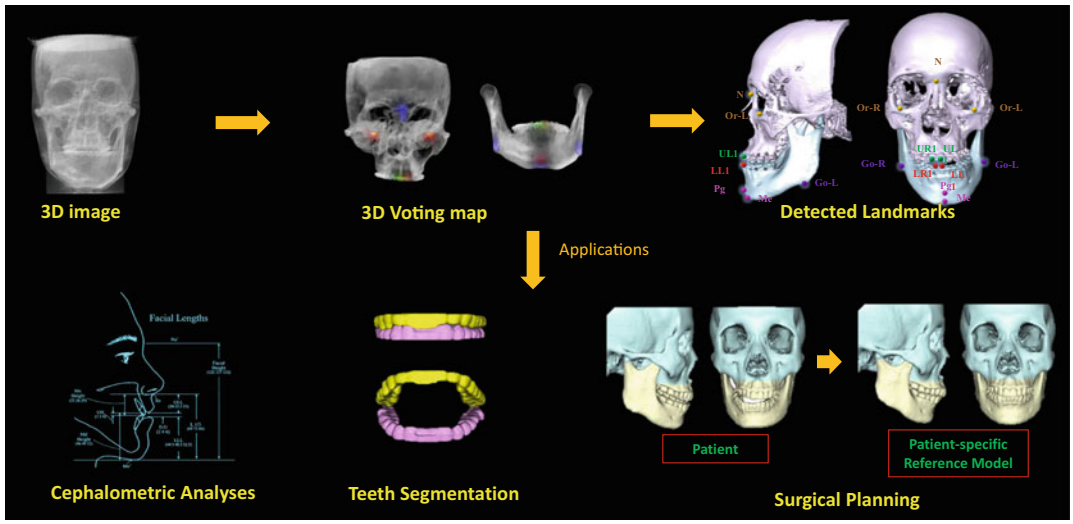


Fig. 2.2 Landmark digitization and its applications

use because of two major challenges. First is morphological variations among different patients. The local morphological appearance of the same landmark could be significantly different across various patients. For example, Fig. 2.3a shows a tooth landmark in relatively different positions for two patients. The second challenge to automatic landmark digitization is image artifacts, such as amalgam fillings and orthodontic wires, bands, and brackets. The left image of Fig. 2.3b shows artifacts from orthodontic braces, while the right image illustrates a patient without them. In this example, the metal brackets made the local patch appearances near the teeth landmarks inconsistent.

To address these issues, various methods have been proposed for improving landmark digitization in order to achieve a clinically acceptable performance [12, 13]. First, landmark localization methods will be discussed in Sect. 2. Then, four methods for landmark digitization will be presented in Sect. 3. Section 4 presents the experimental results using these four methods on a dataset with 107 CT images. Finally, Sect. 5 concludes this chapter.

2.2 Related Methods for Landmark Localization

A number of techniques have been proposed to automatically localize landmarks and anatomical structures for medical applications. In general, there are three mainstreams: (1) interest point detection [14, 15], (2) atlas-based landmark detection [16–19], and (3) machine learning-based landmark detection [20–22].

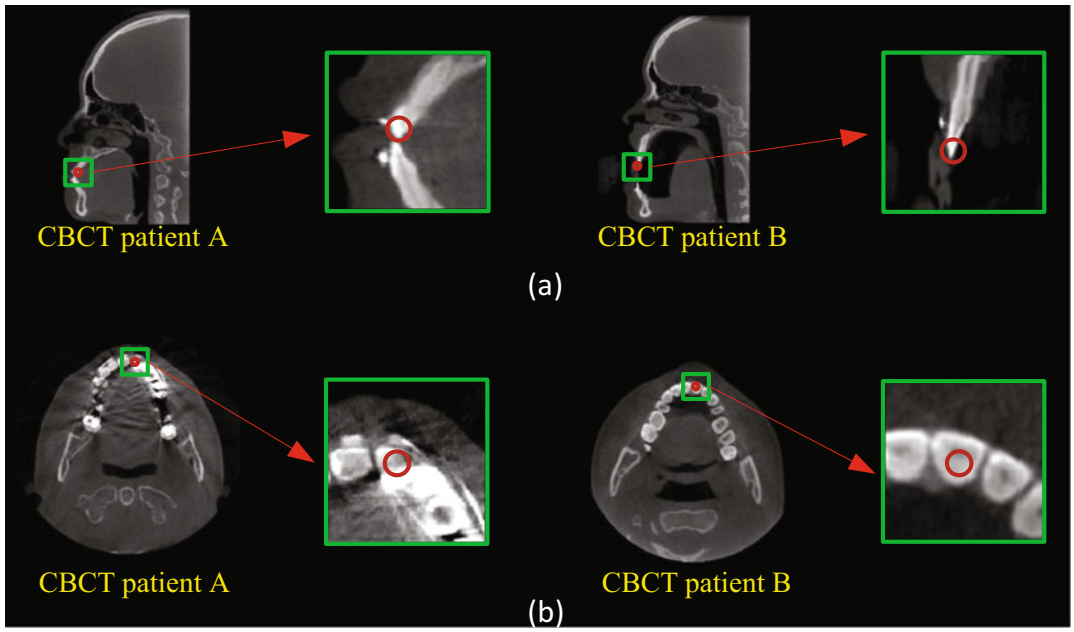


Fig. 2.3 Various morphological appearances. (**a**) Morphological variations among different patients. (**b**) Imaging artifacts of CBCT

Among these options, machine learning-based methods have become increasingly popular in landmark localization. Previous learning-based methods focus on using a voxel-wise classification strategy to localize anatomical landmarks.

Here, the localization is formulated as a binary classification problem, where voxels near the landmark are regarded as positives and the rest as negatives. Then, a classifier is typically trained to distinguish landmark voxels from the others. For example, Zhan et al. [23] detected anatomical landmarks of multiple organs via confidence maximizing sequential scheduling. Criminisi et al. [24] used a classification forest to automatically localize the bounding boxes of multiple organs in CT images. Cheng et al. [25] used a random forest classifier to localize a dent landmark. Zhan et al. [26] used cascade Ada-boost classifier for knee landmark detection with magnetic resonance imaging (MRI) scans. Since classification-based methods rely on only the local appearance for landmark localization, their performance is jeopardized if this appearance is inconsistent across patients.

Regression-based methods are another type of machine learning-based method used for landmark localization. Different from classification-based methods, regression-based methods aim to learn a mapping from a voxel's appearance to its 3D displacement toward a landmark. When provided with a testing image, the 3D displacement from every voxel to the target landmark can be estimated by the learned mapping. Hence, every voxel can cast one vote to the potential landmark position, pointed by the estimated displacement. By aggregating all votes, the landmark position can be localized at the voxel that receives the maximum votes. The regression-based methods can potentially overcome the limitations of classification-based methods by borrowing the context information from nearby voxels with consistent local appearances. Recently, regression forest-based methods have demonstrated their superiority in various computer vision and medical tasks [27–32]. For example, Criminisi et al. [27] proposed to use regression forest to estimate the 3D displacement from each voxel to the bounding box of a target organ. Their experimental results demonstrated that regression-based methods are more accurate than classification-based methods in bounding box detection. With the similar regression voting

strategy, Cootes et al. [28] extended the work of [27] for facial landmark localization. To further enforce spatial consistency of localized landmarks, Gao et al. [32] proposed a two-layer context-aware regression forest for prostate landmark detection. The problem with these methods is that they treat the vote from every voxel equally. As a result, the final localization result can be impacted by noisy votes from uninformative voxels. To address this issue, Donner et al. [33] proposed to use a classifier to pre-filter the voting voxels, by allowing only the voxels near the landmark to vote. Besides using regression forest for estimating the displacement to the target landmark, Chen et al. [34, 35] developed a data-driven method to jointly estimate displacements from all patches to landmarks together, thus achieving remarkable accuracy on X-ray landmark detection as well as on intervertebral disc localization from MR images.

Recent studies have employed an end-to-end learning strategy for anatomical landmark detection, through which the relationship among landmarks and patches can be captured. In these methods, landmark detection is often formulated as a regression problem [36, 37], with the goal of learning a nonlinear mapping between an input image and landmark coordinates. In this way, the landmark position can be directly estimated via deep learning models, such as convolutional neural networks (CNNs). Besides, fully convolutional networks (FCNs) have achieved impressive performance in object detection and landmark detection [38–41]. Recently, Payer et al. [42] tested several FCN architectures for detecting anatomical landmarks with limited medical imaging data, where each landmark position was marked as a heat map corresponding to the original image. Reasonable detection performance was obtained, illustrating the effectiveness of FCN in detecting anatomical landmarks. However, due to the limited number of training data, very shallow networks were used in the experiments, which could not entirely capture the discriminative information in the medical images. Also, it is challenging to simultaneously detect large-scale landmarks in an end-to-end manner, since each landmark corresponds to an output of a 3D heat map. As a result, the existing GPU memory cannot deal with thousands of 3D output maps jointly. In contrast, if multiple landmarks are detected separately, it is cumbersome to train many models, and the underlying correlations among landmarks will be ignored. To address this problem, Zhang et al. [21] proposed a two-stage task-oriented deep learning (T^2DL) method to detect large-scale anatomical landmarks simultaneously in real time, using limited training data. Two cascaded deep CNNs were developed, with each focusing on a specific task.

2.3 Craniomaxillofacial Landmark Digitization

In this section, four methods for landmark digitization will be introduced, including the multi-atlas (MA)-based method, the regression forest (RF)-based method, the segmentation-guided regression forest, and a deep learning model.

2.3.1 Multi-Atlas-Based Landmark Digitization

The multi-atlas-based landmark digitization method employs the technique of image registration. In general, there are atlas images that have the annotated ground-truth landmarks. In the application stage, the test image needs to be registered to each atlas image using linear registration and nonlinear registration successively. Currently, there are many image registration tools that can be used for both linear registration (e.g., FLIRT in FSL [43]) and nonlinear registration (e.g., Demons [44], Syn [45], and deep-learning-based registration [46, 47]). With the affine matrix obtained from linear registration and deformation field obtained from nonlinear registration, the landmarks from atlas images can be propagated to the target test image. By averaging/majority voting the estimated landmark positions from multiple atlases, the landmark position for the test image can be finally calculated.

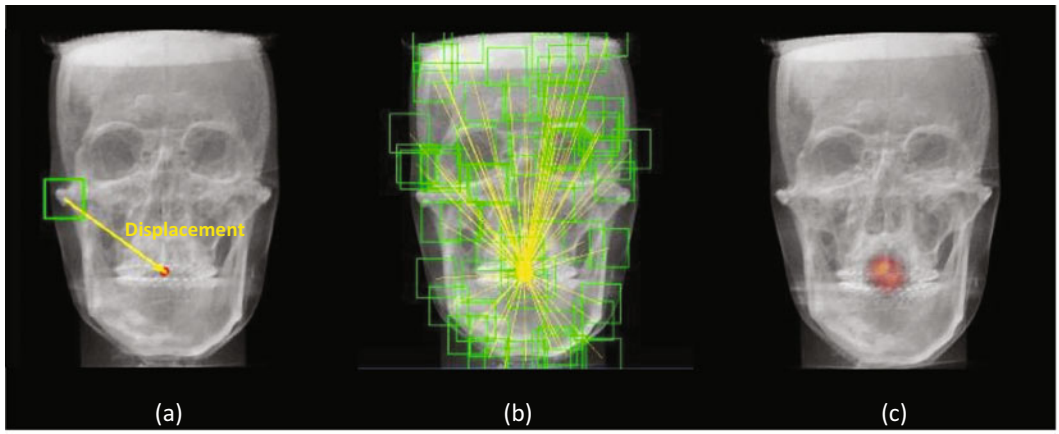


Fig. 2.4 Regression forest for landmark digitization. (**a**) Displacement to a target landmark. (**b**) Majority voting by displacements. (**c**) Heatmap after voting

2.3.2 Regression Forest-Based Landmark Digitization

Regression forest-based methods have demonstrated their superiority in anatomy detection for different organs and structures [28, 48, 30, 29, 32]. Specifically, in the training stage, the regression forest can be used to learn a nonlinear mapping between a voxel's local appearance and its 3D displacement to the target landmark. Generally, Haar-like features are used to describe a voxel's local appearance.

In the testing stage, the learned regression forest can be used to estimate a 3D displacement from every voxel in the image to the potential landmark position, based on local appearance features extracted from the neighborhood of this voxel (see Fig. 2.4a). Using the estimated 3D displacement, each voxel can cast one vote to the potential landmark position (see Fig. 2.4b). By aggregating all votes from all voxels, a voting map can be obtained, from which the landmark position can be easily identified as the location with the maximum vote (see Fig. 2.4c).

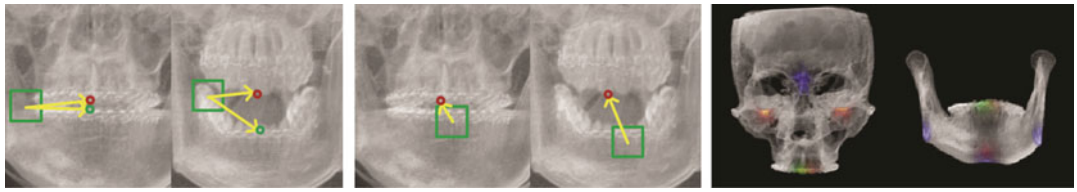
2.3.3 Landmark Digitization Using Segmentation-Guided Partially Joint Regression Forest

Despite its popularity, the traditional regression forest-based methods still suffer from two limitations: (1) all landmarks are often jointly localized without considering the incoherence of landmark positions. Among the CMF landmarks, the landmark positions, respectively, from different anatomical regions could change dramatically because of morphological differences across patients. (2) All image voxels are equally treated in the regression voting stage. Thus, the voting map may be negatively influenced by the voxels that are uninformative to the target landmark position. To address these two limitations, a segmentation-guided partially joint regression forest model is proposed for landmark digitization [13], with details given below.

Partially Joint Regression Forest Model A joint model was developed to predict the displacements of a voxel to multiple landmarks. This model assumes that similar local appearances correspond to coherent multiple displacements.

However, this assumption is invalid in this application. For example, Fig. 2.5a shows two CBCT images, where the displacements to a lower tooth and an upper tooth are not coherent. Since regression forest predicts the displacement based solely on the input appearance features, the ambiguous displacements associated with voxels of same appearance could mislead the training procedure, which could eventually lower the accuracy of the landmark digitization.

To address this issue, the coherence of landmark positions is exploited. Specifically, a partially joint regression forest model is proposed



(a) (b) (c)

Fig. 2.5 Displacement and voting map. (**a**–**b**) Two cases of incoherent displacements happened near teeth landmarks. (**c**) A typical voting map overlaid with 3D rendered skull. Image courtesy to [13]

by dividing all CMF landmarks into several groups. To form groups, landmarks are clustered, based on a dissimilarity matrix, where each entry is the variance of pair-wise landmark distances across subjects. Then, several fully joint models are learned with each model corresponding to a specific group of landmarks. Using this approach, the ambiguous displacement problem can be largely avoided. Besides, in this application, the Haar-like features could be easily affected by the local inhomogeneity caused by artifacts. Therefore, the multi-scale statistical features based on oriented energies [49] are employed in this method. The features are invariant to local inhomogeneity and thus can potentially relieve the effects of artifacts. Instead of using N-ray coding [49] for vector quantization, the bag-of-words strategy is used to achieve relatively low feature dimension. Moreover, multi-scale representation captures both coarse and fine structural information, which can help describe local appearance more effectively.

Segmentation-Guided Strategy As stated above, one limitation of traditional regression forest-based methods is the failure to consider the voting confidence of each voxel and equally treating their votes. Since different voxels have different voting confidences, it is necessary to associate voting weight for each voxel. Here, two types of uninformative voxels are considered.

The first type of uninformative voxels is the faraway voxels. In this case, an intuitive solution is to assign large voting weights only for the voxels near the landmark, while small voting weights are assigned for those faraway voxels, as they are not informative to predict landmark locations. Specifically, a voting weight of a voxel $w = e^{-\frac{\left\|\tilde{\mathrm{d}}\right\|}{\alpha}}$ is defined to reduce the influence of faraway voxels, where $\left\|\tilde{\mathrm{d}}\right\|$ is the estimated 3D displacement from this voxel to the target landmark and α is a scaling coefficient.

The second type of uninformative voxels is voxels that mislead the prediction. In the application, voxels in the mandible often have consistent 3D displacements to the lower teeth landmarks but may have ambiguous 3D displacements to the upper teeth landmarks. This causes the votes from voxels in the lower teeth region to be unreliable for localizing the upper teeth landmarks. To address this issue, image segmentation is used as guidance to remove those uninformative voxels. The original CBCT image is segmented into mandible and maxilla using an automatic segmentation method [1]. Specifically, the deformable registration method was first used to warp all atlases to the current testing image. Then, a sparse representation-based label propagation strategy was employed to estimate a patient-specific atlas from all aligned atlases. Finally, the patient-specific atlas was integrated into a Bayesian framework for accurate segmentation. By using the segmented mandible and maxilla as masks, the mandible can be separated from maxilla in the original CBCT image. Therefore, landmarks on the mandible and maxilla can now be separately digitized using our partially joint model on the respectively masked CBCT images. By using segmentation as guidance, those uninformative voxels can be removed for localizing the landmark.

2.3.4 Joint Bone Segmentation and Landmark Digitization Using Deep Learning

Figure 2.6 illustrates the schematic diagram of the joint bone segmentation and landmark digitization (JSD) framework. A network FCN-1 is developed to learn the *displacement maps* for multiple landmarks from an input image to model the spatial context information in the whole image. The size of each *displacement map* is the same size as the input image, and each element in the displacement map records the displacement from the current voxel location to a respective landmark in a specific axis space. Then, another network FCN-2 is developed to simultaneously perform bone segmentation and landmark digitization by using both the displacement maps estimated by FCN-1 and the original image as the input.

Displacement Map Guided Network Similar to [50], for a 3D image $\mathbf{X}_n$ with V voxels, a *displacement map* is represented as a 3D volume of the same size as $\mathbf{X}_n$, with each element denoting the displacement from a voxel to a certain landmark in a specific axis space. That is, for the l-th landmark in $\mathbf{X}_n$, there are three displacement maps (i.e., $\mathbf{D}^{l,x}{}_n$, $\mathbf{D}^{l,y}{}_n$, and $\mathbf{D}^{l,z}{}_n$) corresponding to x, y, and z axes, respectively. Given L landmarks, 3 L displacement maps are obtained for each input image.

As shown in Fig. 2.7(left), the first subnetwork (i.e., FCN-1) is developed to learn a nonlinear mapping from the input image to the displacement maps. Using a set of training images and their corresponding target displacement maps, FCN-1 adopts a U-net architecture [51] to capture both the global and the local structural information of input images. Specifically, FCN-1 consists of a contracting path and an expanding path. The contracting path follows the typical architecture of a CNN. Every step in the contracting path consists of two $3 \times 3 \times 3$ convolutions, followed by a rectified linear unit (ReLU) and a $2 \times 2 \times 2$ max pooling operation with stride 2 for down sampling. Each step in the expanding path consists of a $3 \times 3 \times 3$ up-convolution, followed by a concatenation with the corresponding feature map from the contracting path, and two $3 \times 3 \times 3$ convolutions (each followed by a ReLU).

Due to the contracting path and the expanding path, such a network is able to grasp a large image area using small kernel sizes while still keeping high localization accuracy. Note that the output of the last layer in FCN-1 is normalized into $[-1,1]$. Let $X_{n,v}$ represent the v-th ($v = 1,\cdots,V$) voxel of the image $\mathbf{X}_n$. In the a-th ($a \in \{x,y,z\}$) axis space, the l-th ($l = 1,\cdots,L$) displacement map of $\mathbf{X}_n$ is denoted as $\mathbf{D}^{l,a}{}_n$, and its v-th element is denoted as $D_{n,v}{}^{l,a}$. The target of FCN-1 is to learn a nonlinear mapping function to transform the original input image onto its corresponding 3 L displacement maps, by minimizing the following loss function:

$$\Omega_1(\mathbf{w_1}) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{n=1}^{N}\frac{1}{V}\sum_{v=1}^{V}\frac{1}{3} \times \sum_{a\in\{x,y,z\}}\left(D_{n,v}^{l,a} - f\left(X_{n,v}, \mathbf{w}_1\right)\right)^2, \tag{1}$$

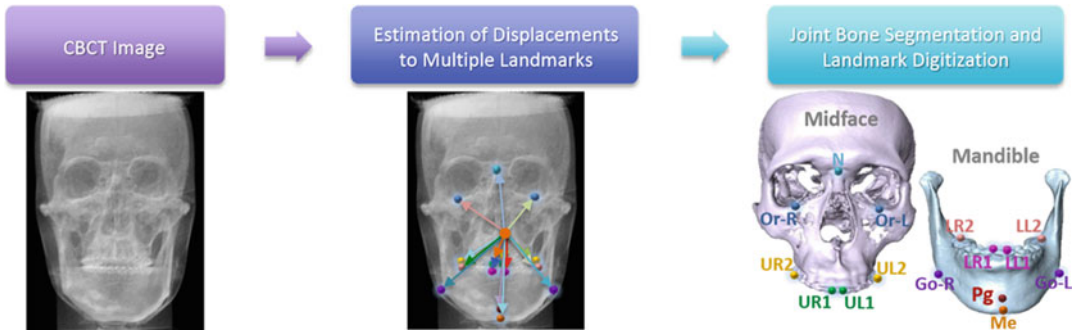


Fig. 2.6 Pipeline of the JSD framework. Image courtesy to [12]

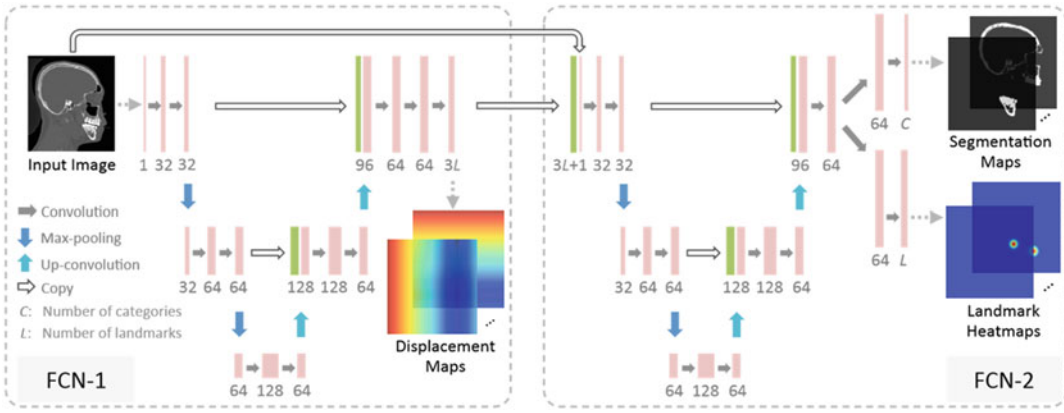


Fig. 2.7 Overview of displacement map guided multi-task FCN. FCN-1 estimates the displacement maps, while FCN-2 performs joint bone segmentation and landmark digitization. Image courtesy to [12]

where $f(X_{n,v},\mathbf{W}_1)$ is the estimated displacement by using the network coefficients $\mathbf{W}_1$ and N is the number of training images in a batch.

Joint Bone Segmentation and Landmark Digitization Using the displacement maps learned in FCN-1 as guidance, a subnetwork (i.e., FCN-2) with a U-net architecture is further designed to jointly perform bone segmentation and landmark digitization. As shown in Fig. 2.7(right), FCN-2 uses a stacked representation of displacement maps and the original image as the input, through which the spatial context information is explicitly incorporated into the learning process. In addition, such representation could guide the network to focus on the informative regions and thus may help alleviate the negative influence of image artifacts. Note that the output of the last layer for bone segmentation is transformed to probability scores by using the softmax function, and that for landmark digitization are normalized to [0,1]. Denote $\mathbf{Y}_n{}^c$ as the ground-truth segmentation map for the c-th ($c = 1,\cdots,C$) category, with the v-th element as $Y^c_{n,v}$. Here, a CT image is segmented into $C = 3$ categories (i.e., midface, mandible, and background). The ground-truth landmark heatmap for the l-th ($l = 1,\cdots,L$) landmark in $\mathbf{X}_n$ is denoted as $\mathbf{A}^l{}_n$, with its v-th element as $A^l{}_{n,v}$. The objective of FCN-2 is to minimize the following loss function:

$$\begin{aligned}\Omega_2(\mathbf{w}_2) = &-\tfrac{1}{C}\sum_{c=1}^{C}\tfrac{1}{N}\sum_{n=1}^{N}\tfrac{1}{V}\sum_{v=1}^{V}1\left\{Y^c_{n,v}=c\right\}\\ &\times\log\left(\mathbf{P}\left(Y^c_{n,v}=c|\,X_{n,v};\mathbf{w}_2\right)\right)\\ &+\tfrac{1}{L}\sum_{l=1}^{L}\tfrac{1}{N}\sum_{n=1}^{N}\tfrac{1}{V}\\ &\times\sum_{v=1}^{V}\left(A^l_{n,v}-g\left(X_{n,v},\mathbf{w}_2\right)\right),\end{aligned} \tag{2}$$

where the first term is the cross-entropy error for bone segmentation and the second term is the mean squared error for landmark digitization. Here, $1\{\cdot\}$ is an indicator function, with $1\{\cdot\} = 1$ if $\{\cdot\}$ is true; and 0, otherwise. $\mathbf{P}\left(Y^c_{n,v} = c|\,X_{n,v};\mathbf{w}_2\right)$ indicates the probability of the v-th voxel in the image $\mathbf{X}_n$ being correctly classified as the category $Y^c_{n,v}$ using the network coefficients $\mathbf{W}_2$. The second term in Eq. (2) computes the loss between the estimated value $g(X_{n,v},\mathbf{W}_2)$ and the ground-truth $A^l_{n,v}$ in the l-th landmark heatmap.

Learning Strategy For each landmark, a heatmap is generated using a Gaussian filtering with the standard derivation of 2 *mm* and then stretching the values to the range of [0,1].

For optimizing the network coefficients, the stochastic gradient descent (SGD) algorithm combined with the backpropagation algorithm is adopted. In the training stage, network FCN-1 is first trained using the CT image and its corresponding target displacement maps as the input and the output, respectively.

With FCN-1 frozen, network FCN-2 is further trained for joint bone segmentation and landmark digitization, by using the stacked representation of the estimated displacement maps from FNC-1 and the original image as the input, while landmark heatmaps and segmentation maps served as the output. Finally, using the learned coefficients as initialization, networks FCN-1 and FCN-2 are trained jointly. In addition, the training process is done in a sliding window fashion (with the fixed window size of $96 \times 96 \times 96$). Then, a new testing image of any size can be fed into the trained model, since FCN only contains the convolutional computation.

2.4 Results

Table 2.1 shows the mean error of landmark digitization of four methods. Obviously, the deep learning-based method (i.e., JSD) achieves the best performance of 1.10 *mm*. The special designed random forest-based method (i.e., SPRF) also achieved a competitive result of 1.52 *mm*.

Table 2.1 Digitization errors of four methods

Method	MA RF SPRS JSD
Error	3.05 *mm* 2.67 *mm* 1.52 *mm* 1.10 *mm*

Specifically, Fig. 2.8 shows the digitization errors for each of 15 anatomical landmarks achieved by 4 different methods. From Fig. 2.8, we can see that, compared with MA and RF, SPRF and JSD generally achieve the lower errors in detecting these 15 landmarks, especially for the landmarks located at the lower teeth and upper teeth (e.g., LR1, LL1, LR2, LL2, UR1, UL1, UR2, and UL2, see Fig. 2.6). It is worth noting that, because of large inter-subject variations in the local appearance of the teeth, it is very challenging to accurately localize tooth landmarks. These results demonstrate that our proposed two strategies (i.e., using displacement maps as guidance information and joint learning) can help accurately locate anatomical landmarks in CBCT images. Also, it is clinically acceptable if the digitization error of CMF landmarks for CBCT images is below 1.50 *mm*. Table 2.1 and Fig. 2.8 clearly demonstrate that the average digitization error achieved by our JSD method is below 1.50 *mm*, indicating that JSD has a great value in real clinical applications. Figure 2.9 visually shows the landmark digitization results for three patients using four different methods.

2.5 Conclusion

In this chapter, several landmark localization methods were reviewed, and their potential applications to CMF landmark digitization were discussed. Four typical methods were described in detail. The experimental results on 107 images demonstrated the effectiveness of current landmark digitization methods. The landmark

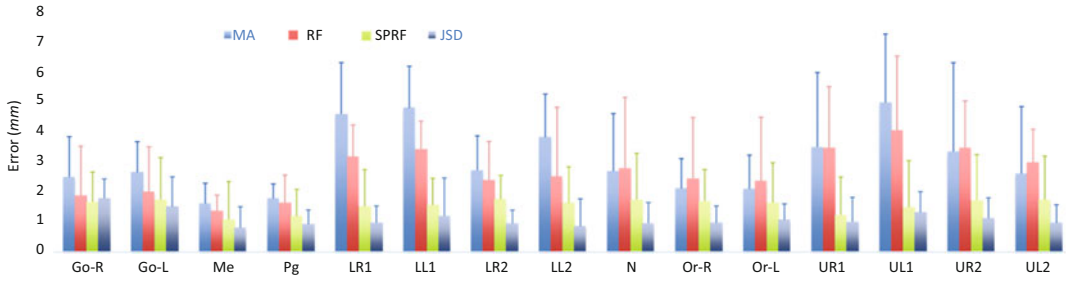


Fig. 2.8 The digitization errors (*mm*) for each of 15 landmarks, achieved by 4 different methods. Image courtesy to [12]

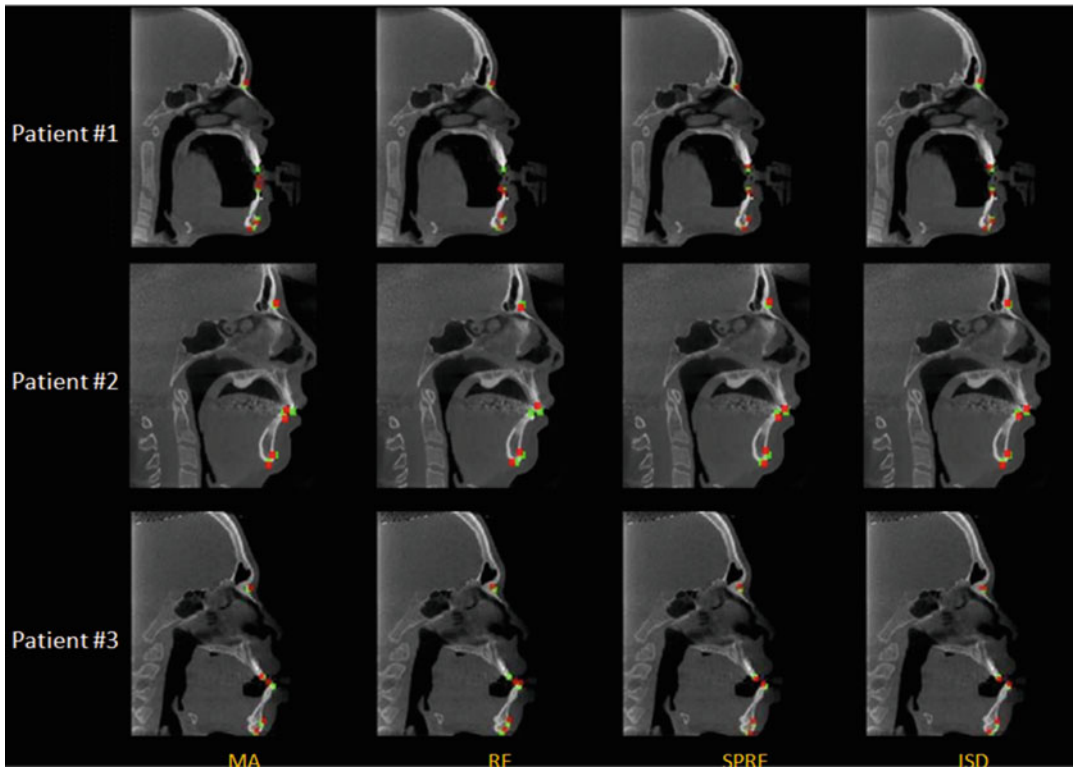


Fig. 2.9 Landmark digitization results for three patients using four different methods. The green points are the ground-truth landmarks, and the red points are the detected landmarks. Image courtesy to [52]

digitization error is clinically acceptable for the methods of SPRF and JSD. Although the deep learning-based method (i.e., JSD) has achieved promising results, there are still several limitations. First, there are only 107 images at hand for model learning. It is possible to augment the training images by using synthetic data (e.g., by using deformable transformation or Generative Adversarial Networks) to further improve the robustness of the method. Second, the tasks of bone segmentation and landmark digitization are treated equally, without considering their possible different contributions. A reasonable solution is to automatically learn the optimal weights for these different tasks from data.

References

1. Wang L, Chen KC, Gao Y, Shi F, Liao S, Li G, Shen SG, Yan J, Lee PK, Chow B, et al. Automated bone segmentation from dental cbct images using patch-based sparse representation and convex optimization. Med Phys. 2014;41(4):043503.
2. Li Z, An L, Zhang J, Wang L, Xia JJ, Shen D. Craniomaxillofacial deformity correction via sparse representation in coherent space. In: International workshop on machine learning in medical imaging: Springer; 2015. p. 69–76.
3. Zhang J, Cain EH, Saha A, Zhu Z, Mazurowski MA. Breast mass detection in mammography and tomosynthesis via fully convolutional network-based heatmap regression. In: Medical imaging 2018: computer-aided diagnosis, vol. 10575: International Society for Optics and Photonics; 2018. p. 1057525.

4. Zhang J, Saha A, Zhu Z, Mazurowski MA. Hierarchical convolutional neural networks for segmentation of breast tumors in mri with application to radiogenomics. IEEE Trans Med Imaging. 2018;38(2):435–47.
5. Cao X, Yang J, Zhang J, Wang Q, Yap PT, Shen D. Deformable image registration using a cue-aware deep regression network. IEEE Trans Biomed Eng. 2018;65(9):1900–11.
6. Liu M, Zhang J, Nie D, Yap PT, Shen D. Anatomical landmark based deep feature representation for mr images in brain disease diagnosis. IEEE J Biomed Health Inform. 2018;22(5):1476–85.
7. Lian C, Liu M, Zhang J, Shen D. Hierarchical fully convolutional network for joint atrophy localization and alzheimer's disease diagnosis using structural mri. IEEE Trans Pattern Anal Mach Intell. 2018; https://doi.org/10.1109/TPAMI.2018.2889096.
8. Liu M, Zhang J, Adeli E, Shen D. Landmark-based deep multi-instance learning for brain disease diagnosis. Med Image Anal. 2018;43:157–68.
9. Zhang J, Liu M, An L, Gao Y, Shen D. Alzheimer's disease diagnosis using landmark-based features from longitudinal structural mr images. IEEE J Biomed Health Inform. 2017;21(6):1607–16.
10. Liu M, Zhang J, Adeli E, Shen D. Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis. IEEE Trans Biomed Eng. 2018;66(5):1195–206.
11. Liu M, Zhang J, Lian C, Shen D. Weakly supervised deep learning for brain disease prognosis using mri and incomplete clinical scores. IEEE transactions on cybernetics. 2019;
12. Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen SGF, Tang Z, Chen KC, Xia JJ, et al. Joint craniomaxillofacial bone segmentation and landmark digitization by context-guided fully convolutional networks. In: International conference on medical image computing and computer-assisted intervention: Springer; 2017. p. 720–8.
13. Zhang J, Gao Y, Wang L, Tang Z, Xia JJ, Shen D. Automatic craniomaxillofacial landmark digitization via segmentation-guided partially-joint regression forest model. In: International conference on medical image computing and computer-assisted intervention: Springer; 2015. p. 661–8.
14. Donner R, Micuvsik B, Langs G, Bischof H. Sparse mrf appearance models for fast anatomical structure localisation. In: Proc: BMVC; 2007.
15. Donner R, Langs G, Mivcuvsik B, Bischof H. Generalized sparse mrf appearance models. Image Vis Comput. 2010;28(6):1031–8.
16. Nowinski WL, Thirunavuukarasuu A. Atlas-assisted localization analysis of functional images. Med Image Anal. 2001;5(3):207–20.
17. Yelnik J, Damier P, Demeret S, Gervais D, Bardinet E, Bejjani BP, Franccois C, Houeto JL, Arnulf I, Dormont D, et al. Localization of stimulating electrodes in patients with parkinson disease by using a three-dimensional atlas-magnetic resonance imaging coregistration method. J Neurosurg. 2003;99(1):89–99.
18. Fenchel M, Thesen S, Schilling A. Automatic labeling of anatomical structures in mr fastview images using a statistical atlas. In: Medical image computing and computer-assisted intervention–MICCAI 2008: Springer; 2008. p. 576–84.
19. Shahidi S, Bahrampour E, Soltanimehr E, Zamani A, Oshagh M, Moattari M, Mehdizadeh A. The accuracy of a designed software for automated localization of craniofacial landmarks on cbct images. BMC Med Imaging. 2014;14(1):32.
20. Zhang J, Gao Y, Wang L, Tang Z, Xia JJ, Shen D. Automatic craniomaxillofacial landmark digitization via segmentation-guided partially-joint regression forest model and multiscale statistical features. IEEE Trans Biomed Eng. 2015;63(9):1820–9.
21. Zhang J, Liu M, Shen D. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans Image Process. 2017;26(10):4753–64.
22. Zhang J, Gao Y, Gao Y, Munsell B, Shen D. Detecting anatomical landmarks for fast Alzheimer's disease diagnosis. IEEE Trans Med Imaging. 2016;35(12):2524–33.
23. Zhan Y, Zhou XS, Peng Z, Krishnan A. Active scheduling of organ detection and segmentation in whole-body medical images. In: Medical image computing and computer-assisted intervention–MICCAI 2008: Springer; 2008. p. 313–21.
24. Criminisi A, Shotton J, Bucciarelli S. Decision forests with long-range spatial context for organ localization in ct volumes. In: Medical image computing and computer-assisted Interventation (MICCAI); 2009. p. 69–80.
25. Cheng E, Chen J, Yang J, Deng H, Wu Y, Megalooikonomou V, Gable B, Ling H. Automatic dent-landmark detection in 3-d cbct dental volumes. In: Engineering in medicine and biology society, EMBC, 2011: Annual International Conference of the IEEE; 2011. p. 6204–7.
26. Zhan Y, Dewan M, Harder M, Krishnan A, Zhou XS. Robust automatic knee MR slice positioning through redundant and hierarchical anatomy detection. IEEE Tran on Medical Imaging. 2011;30(12):2087–100.
27. Criminisi A, Shotton J, Robertson D, Konukoglu E. Regression forests for efficient anatomy detection and localization in ct studies. In: Medical computer vision. Recognition techniques and applications in medical imaging: Springer; 2011. p. 106–17.
28. Cootes TF, Ionita MC, Lindner C, Sauer P. Robust and accurate shape model fitting using random forest regression voting. In: ECCV 2012: Springer; 2012. p. 278–91.
29. Criminisi A, Robertson D, Konukoglu E, Shotton J, Pathak S, White S, Siddiqui K. Regression forests for efficient anatomy detection and local-

ization in computed tomography scans. Med Image Anal. 2013;17(8):1293–303.
30. Lindner C, Thiagarajah S, Wilkinson JM, Consortium T, Wallis G, Cootes T. Fully automatic segmentation of the proximal femur using random forest regression voting. Medical Imaging, IEEE Transactions on. 2013;32(8):1462–72.
31. Chu C, Chen C, Wang CW, Huang CT, Li CH, Nolte LP, Zheng G. Fully automatic cephalometric x-ray landmark detection using random forest regression and sparse shape composition. submitted to Automatic Cephalometric X-ray Landmark Detection Challenge. 2014.
32. Gao Y, Shen D. Context-aware anatomical landmark detection: application to deformable model initialization in prostate ct images. In: Machine learning in medical imaging: Springer; 2014. p. 165–73.
33. Donner R, Menze BH, Bischof H, Langs G. Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization. Med Image Anal. 2013;17(8):1304–14.
34. Chen C, Xie W, Franke J, Grutzner P, Nolte LP, Zheng G. Automatic x-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. Med Image Anal. 2014;18(3):487–99.
35. Chen, C., Belavy, D., Yu, W., Chu, C., Armbrecht, G., Bansmann, M., Felsenberg, D., Zheng, G.: Localization and segmentation of 3d intervertebral discs in mr images by data driven estimation. (2015).
36. Zheng Y, Liu D, Georgescu B, Nguyen H, Comaniciu D. 3D deep learning for efficient and robust landmark detection in volumetric data. In: International conference on medical image computing and computer-assisted intervention: Springer; 2015. p. 565–72.
37. Riegler G, Urschler M, Ruther M, Bischof H, Stern D. Anatomical landmark detection in medical applications driven by synthetic data. In: Proceedings of the IEEE international conference on computer vision workshops; 2015. p. 12–6.
38. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv. 2013;1312:6229.
39. Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C. Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 648–56.
40. Liang Z, Ding S, Lin L. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. arXiv preprint arXiv. 2015;1507:03409.
41. Dai J, Li Y, He K, Sun J. R-FCN: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems; 2016. p. 379–87.
42. Payer C, Štern D, Bischof H, Urschler M. Regressing heatmaps for multiple landmark localization using CNNs. In: MICCAI: Springer; 2016. p. 230–8.
43. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. Fsl Neuroimage. 2012;62(2):782–90.
44. Thirion JP. Image matching as a diffusion process: an analogy with maxwell's demons. Med Image Anal. 1998;2(3):243–60.
45. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal. 2008;12(1):26–41.
46. Cao X, Yang J, Zhang J, Nie D, Kim M, Wang Q, Shen D. Deformable image registration based on similarity-steered cnn regression. In: International conference on medical image computing and computer-assisted intervention: Springer; 2017. p. 300–8.
47. Zhang J. Inverse-consistent deep networks for unsupervised deformable image registration. arXiv preprint arXiv. 2018;1809:03443.
48. Zhang S, Zhan Y, Dewan M, Huang J, Metaxas DN, Zhou XS. Towards robust and effective shape modeling: sparse shape composition. Med Image Anal. 2012;16(1):265–77.
49. Zhang J, Liang J, Zhao H. Local energy pattern for texture classification using self-adaptive quantization thresholds. Image Processing, IEEE Transactions on. 2013;22(1):31–42.
50. Pfister T, Charles J, Zisserman A. Flowing convnets for human pose estimation in videos. In: ICCV; 2015. p. 1913–21.
51. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: MICCAI: Springer; 2015. p. 234–41.
52. Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen SG, Tang Z, Chen KC, Xia JJ, Shen D. Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. Med Image Anal. 2020;60:101621.

Segmenting Bones from Brain MRI via Generative Adversarial Learning

3

Xu Chen, Chunfeng Lian, Li Wang, Pew-Thian Yap, James J. Xia, and Dinggang Shen

3.1 Introduction

Accurate segmentation of bony structures to construct high-quality skeletal models is important for diagnosis and planning the treatment of craniomaxillofacial (CMF) deformities. Computed tomography (CT) is commonly used in clinical practice for skeletal model construction due to its high contrast between bony structures and surrounding tissues. However, since CT scanning emits ionizing radiation that is harmful to the human body (especially for infant patients), finding an alternative imaging modality for safer bony anatomy acquirement is of great clinical value. Compared with CT, magnetic resonance imaging (MRI) is non-ionizing and much safer. The major challenge for MRI-based CMF anatomy assessment lies in the segmentation of bony structures (even for experienced radiologists), considering that they typically lack contrast and are indistinguishable from the air in MRI images (see Fig. 3.1). Thus, the manual bony annotations are typically unavailable for CMF MRI scans, which is a major hurdle in building a reliable bony structure segmentation model for MRI.

Recently, several deep learning methods have been proposed to automatically segment CMF bony structures from brain MRI. To tackle the above challenge associated with MRI images, these methods usually integrate supplementary bony information from the paired CT images (i.e., captured from the same subjects) to assist the training of segmentation models on MRI data, since the anatomical structures (bony structures) of paired MRI-CT images are considered to be consistent. For example, Nie et al. [1] proposed a cascaded framework to segment MRI bony structures in a coarse-to-fine fashion, in which annotations from corresponding CT images were used as the ground truth for model construction. Similarly, using training samples with paired MRI-CT, Zhao et al. [2] proposed to first learn a

X. Chen (✉) · C. Lian · L. Wang · P.-T. Yap · D. Shen
BRIC and Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: chenxu31@email.unc.edu; dgshen@med.unc.edu

J. J. Xia
Department of Oral and Maxillofacial Surgery, Houston Methodist Research Institute, TX, Houston, USA

Department of Surgery (Oral and Maxillofacial Surgery), Weill Medical College, Cornell University, New York, NY, USA
e-mail: jxia@houstonmethodist.org

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_3

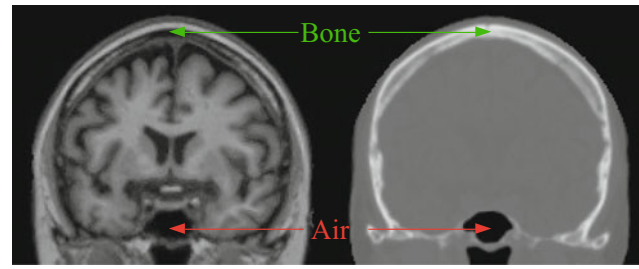


Fig. 3.1 An example of paired MRI-CT images. Compared with CT, the bones and air are both dark and indistinguishable in MRI

generative adversarial network (GAN) to estimate realistic CT images from the MRI images and then perform bony structure segmentation using both MRI images and synthesized CT images. While these methods have achieved relatively good segmentation results, their performance (especially generalization ability) is still hampered by a practical challenge. That is, patients with CMF deformities routinely only take CT scans, which means the number of paired MRI-CT images is extremely limited. On the other hand, important bony information from unpaired CT images is ignored by these methods due to their requirement of paired training images.

Inspired by the recent studies on image-to-image translation [3, 4], in this work, we propose to ameliorate the annotation limitation problem by transferring the bony annotation information from CT modality to MRI modality via cross-modality image synthesis. Specifically, we propose to make full use of both paired and unpaired MRI-CT data to construct a semi-supervised deep neural network for automated CMF bony structure segmentation from MRI. Our proposed network consists of a CycleGAN [3]-based cross-modality image synthesis sub-network and a segmentation sub-network. The image synthesis sub-network learns the cycle-consistent mappings between MRI and CT modalities. Then the synthetic MRI images, which are assumed to contain the same anatomical structure information as their CT inputs, can be used to train the segmentation sub-network by "borrowing" the annotation information (i.e., ground truth) from CT images. However, in contrast to the original CycleGAN that uses only unpaired images for training, we train our image synthesis sub-network with both paired and unpaired MRI-CT data in a semi-supervised manner. To alleviate the intrinsic ambiguity problem of CycleGAN, we propose a novel *neighbor-based anchoring method* to narrow down the space of possible mappings in cross-modality image synthesis by making full use of the paired MRI-CT data. Moreover, a *feature-matching-based semantic consistency constraint* is applied to encourage the image synthesis sub-network to preserve the high-level semantic information (i.e., the anatomical structure information) during cross-modality synthesis. Both qualitative and quantitative experimental results demonstrate the effectiveness of the proposed method in effectively improving the segmentation performance, compared with the state-of-the-art medical image segmentation methods.

To summarize, the main contribution of this work is threefold:

- We propose a generative adversarial learning framework to address the label limitation problem in MRI segmentation of CMF bony structures. This is enabled by transferring the labels from CT modality to MRI modality via cross-modality image synthesis. To be specific, the proposed framework contains a cycle-consistent image synthesis sub-network to learn the cross-modality mappings between MRI and CT modalities using both paired and unpaired MRI-CT data, and the synthetic MRI images are then used to augment the training dataset along with the bony annotations from CT images to guide the training of segmentation sub-network.

- To reduce the ambiguity problem in cycle-consistent cross-modality image synthesis, we propose a novel neighbor-based anchoring method to effectively reduce the space of possible mappings in synthesis by making full use of the paired MRI-CT data.
- A feature-matching-based semantic consistency constraint is applied to regularize the cross-modality image synthesis by encouraging the high-level semantic consistency between the synthetic images and their corresponding inputs.

3.2 Related Works

3.2.1 Semi-supervised Learning

Depending on whether the label is used in the training stage, the machine learning techniques can be roughly categorized into three groups: the supervised learning (using completely labeled data) [5, 6], the unsupervised learning (using only unlabeled data) [7, 8], and the semi-supervised learning (using both labeled and unlabeled data) [9, 10]. The semi-supervised learning aims at improving the generalization abilities of machine learning models by making use of the unlabeled data to train the model. It has found a wide range of applications in many cognitive tasks where the labeled data is limited. For example, Bennett et al. [9] introduced a semi-supervised support vector machine method to address the transduction problem by using a mixture of labeled data and unlabeled data. Weston et al. [10] proposed a semi-supervised framework to improve supervised learning for deep architectures by jointly learning an embedding task on unlabeled data. In particularly, the semi-supervised learning can be of great practical value in medical image segmentation [11, 12], since manually annotating medical images (especially for 3D medical images) is typically tedious and requires professional knowledge and skills. For example, Bai et al. [11] proposed a semi-supervised segmentation network for cardiac MRI image segmentation, where the model was trained on both labeled and unlabeled data. Mahapatra et al. [12] proposed a novel strategy for segmenting Crohns disease tissues from abdominal MRI images by using semi-supervised learning to predict the missing labels. In this work, we propose a semi-supervised framework to build an automatic MRI-based bony structure segmentation network by transferring the annotation information from CT modality to MRI modality via cross-modality image synthesis, where both paired and unpaired MRI-CT data are used to improve the generalizability of the segmentation model.

3.2.2 Cross-Modality Image Synthesis

In clinical practice, multi-modality imaging is often used to contribute the comprehensive diagnosis of complex disease, since different imaging modalities provide complementary information about the patients from respective aspect. However, multi-modality imaging scans are not always available due to many practical factors, such as the scanner unavailability, the time-consuming scanning process, and the cost. Thus, it would be clinically helpful if the desirable target modality image can be generated from the source modality one via cross-modality image synthesis. Earlier efforts on cross-modality image synthesis typically require paired images from two different imaging modalities to learn the translation under some sort of regularizations [13, 14]. For example, Huang et al. [13] proposed a geometry regularized joint dictionary learning framework for MRI cross-modality synthesis. Recently, the studies on unsupervised image-to-image translation have attracted a great deal of interest [3, 4]. Inspired by these works, many researchers proposed to address the unsupervised cross-modality image synthesis problem by using cycle-consistent regularization [15, 16]. For example, Hiasa et al. [15] addressed the unsupervised MRI-to-CT synthesis by extending the CycleGAN method with a gradient consistency loss to improve the accuracy at the boundaries. Huang et al. [16] proposed a weakly supervised framework to jointly solve the problems of super-resolution

and cross-modality image synthesis, aimed at generating high-resolution and multimodal images from low-resolution single-modality imagery. In this work, we propose a semi-supervised framework to transfer the annotation information from a label-rich imaging modality (i.e., CT modality) to a label-poor imaging modality (i.e., MRI modality) via cross-modality image synthesis, then the synthetic images are used to assist in building a reliable segmentation model for the label-poor imaging modality.

3.3 Method

In this section, we introduce the proposed method in detail. We first present an overview of the proposed framework in Sect. 3.3.1. Then we elaborate the two main components of the proposed framework, i.e., the cross-modality image synthesis sub-network and the segmentation sub-network, in Sect. 3.3.2. The novel neighbor-based anchoring method and the feature-matching-based semantic consistency constraint are introduced in Sect. 3.3.3 and Sect. 3.3.4, respectively. Subsequently, we elaborate the detailed structures of all sub-networks in Sect. 3.3.5. Finally, the implemental details are presented in Sect. 3.3.6.

3.3.1 Overview

The schematic diagram of our proposed semi-supervised framework is shown in Fig. 3.2, which consists of (1) a cross-modality image synthesis sub-network and (2) a segmentation sub-network for MRI bony structure annotation. The image synthesis sub-network learns the cross-modality mappings from both MRI-to-CT and CT-to-MRI with two generators $G_{\mathrm{M}\to\mathrm{C}}$ and $G_{\mathrm{C}\to\mathrm{M}}$, respectively. Both paired and unpaired MRI-CT images are used to train the image synthesis sub-network. To guide the training of unsupervised cross-modality image synthesis (i.e., using unpaired images), we adopt the cycle-consistent constraint as a regularization, which requires two generators to be the inverse mappings of each other. Moreover, two discriminators D_{M} and D_{C} are used to encourage the generators to synthesize realistic images via adversarial learning. Through the cross-modality image synthesis, the annotation information in CT modality can be implicitly transferred to the MRI modality, as the synthetic images are expected to contain the same anatomical structures as their corresponding inputs. Thus, we can use the synthetic MRI images to train the segmentation sub-network by "borrowing" the annotations from their corresponding inputs (i.e., CT images). The details of the proposed framework will be introduced in Sect. 3.3.2.

Notably, our model is a semi-supervised framework which differs from the original CycleGAN-based methods. Our model is uniquely trained on both paired and unpaired images to effectively alleviate the ambiguity problem of CycleGAN (i.e., there exits numerous feasible mappings that satisfy the cycle-consistent requirement). To make full use of the paired images, we propose a novel neighbor-based anchoring method to reduce the ambiguity problem in cross-modality image synthesis for unpaired images. In addition, a feature-matching-based semantic consistency constraint is proposed to effectively preserve the anatomical structure information in image synthesis, which is critical to the subsequent segmentation task. These two key components will be introduced in Sect. 3.3.3 and Sect. 3.3.4, respectively.

Finally, we elaborate the structure of each sub-network and the implementation details in Sect. 3.3.5 and Sect. 3.3.6, respectively.

3.3.2 Cross-Modality Image Synthesis for Segmentation

In this work, we propose to build a reliable segmentation model for MRI modality by making use of the annotation information from CT modality. This is enabled by transferring the annotation information from CT to MRI via cross-modality image synthesis. Without loss of generality, denoted by $\left(\mathcal{M}_p, \mathcal{C}_p\right)$ the paired MRI-CT datasets in which the MRI-CT images are one-to-one cor-

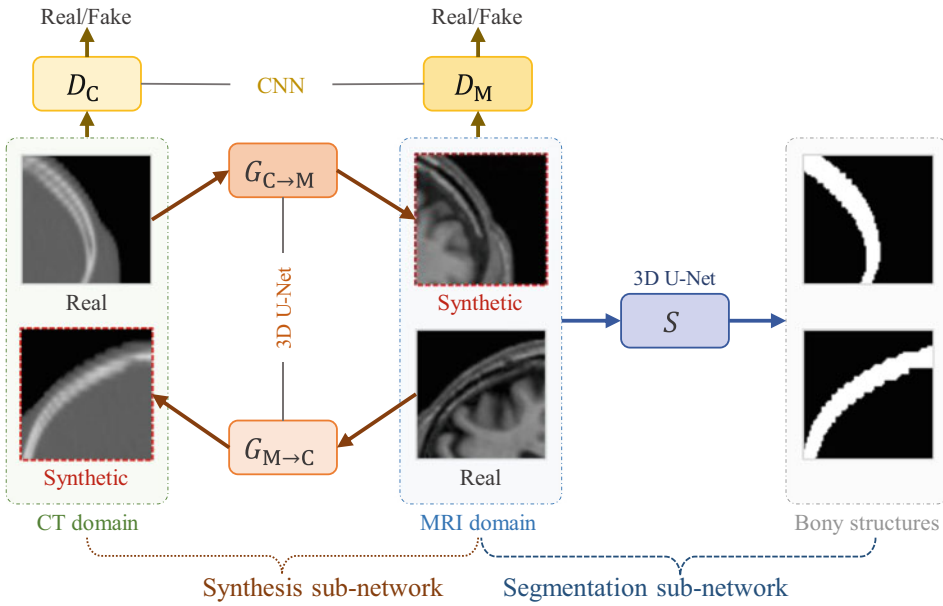


Fig. 3.2 Method overview. Two generators learn the mappings between MRI and CT modalities, two discriminators distinguish the real images from the synthetic ones, and a segmentor estimates bony structures from MRI images. The synthetic MRI data serves as supplementary training data for the segmentation sub-network

respondence (captured from the same subject), and $(\mathcal{M}_u, \mathcal{C}_u)$ the unpaired MRI-CT datasets. In our framework illustrated in Fig. 3.2, the generators $G_{\mathrm{C}\to\mathrm{M}}$ and $G_{\mathrm{M}\to\mathrm{C}}$ learn the cycle-consistent cross-modality mappings between the MRI and CT modalities using both paired and unpaired MRI-CT data. Due to the GPU memory limitation, our model is actually trained on small 3D patches. That is, the datasets $(\mathcal{M}_p, \mathcal{M}_u)$ and $(\mathcal{C}_p, \mathcal{C}_u)$ are comprised of small patches cropped from the original MRI and CT images, respectively. Naturally, the paired data can directly be used to supervise the training of generators by minimizing the following loss function:

$$L_{pair} = \mathbb{E}_{(x,y)\sim(\mathcal{C}_p,\mathcal{M}_p)} \left(\| G_{C\to M}(x) - y \|_1 + \| G_{M\to C}(y) - x \|_1 \right) \tag{3.1}$$

where $x \in \mathcal{C}_p$ and $y \in \mathcal{M}_p$ denote a pair of CT-MRI patches. For unpaired, we adopt the cycle-consistent constraint as a regularization, with the loss function be defined as follows:

$$L_{unpair} = \mathbb{E}_{(x,y)\sim(\mathcal{C}_u,\mathcal{M}_u)} \left(\| G_{M\to C}\left(G_{C\to M}(x)\right) - x \|_1 + \| G_{C\to M}\left(G_{M\to C}(y)\right) - y \|_1 \right) \tag{3.2}$$

where $x \in \mathcal{C}_u$ and $y \in \mathcal{M}_u$ denote the unpaired CT and MRI patches, respectively.

Two discriminators, D_{M} and D_{C}, are used to encourage the generators to synthesize realistic images via adversarial learning. To be specific, the discriminators and the generators are trained alternately in an adversarial manner. The discriminators are first trained to distinguish the real images from the synthetic ones by minimizing the following loss function:

$$L_{dis} = \mathbb{E}_{(x,y)\sim(\mathcal{C},\mathcal{M})} \left[(D_C(x) - 1)^2 + (D_M(y) - 1)^2 + D_M(G_{C\to M}(x))^2 + D_C(G_{M\to C}(y))^2 \right] \tag{3.3}$$

where $\mathcal{C}$ and $\mathcal{M}$ are the union sets of paired and unpaired datasets for CT and MRI (i.e., $\mathcal{C} = \{\mathcal{C}_p, \mathcal{C}_u\}$, $\mathcal{M} = \{\mathcal{M}_p, \mathcal{M}_u\}$), respectively. Then the generators are encouraged to synthesize

realistic images to fool the discriminators using the following loss function:

$$L_{adv} = \mathbb{E}_{(x,y)\sim(\mathcal{C},\mathcal{M})}\left[\left(D_M\left(G_{C\to M}(x)\right)-1\right)^2 + \left(D_C\left(G_{M\to C}(y)\right)-1\right)^2\right] \tag{3.4}$$

Note that we adopt the least square GAN [17] instead of the original GAN [18] to avoid the training instability issue.

Subsequently, we use both real and synthetic MRI images to train the segmentation sub-network S. To be specific, the real MRI images from the paired dataset can share the annotations (i.e., labels) with their corresponding CT images and thus can directly be used to train the segmentor. We use the cross entropy to measure the segmentation error, and the corresponding loss function is defined as follows:

$$L_{\text{seg}}^{\text{r}} = \mathbb{E}_{(x,b)\sim(\mathcal{M}_p,\mathcal{B}_p)}\left[-b\,logS(x) - (1-b)\log\left(1-S(x)\right)\right] \tag{3.5}$$

where $x \in \mathcal{M}_p$ and $b \in \mathcal{B}_p$ represent an MRI patch from the paired dataset and its corresponding annotation, respectively. On the other hand, the MRI images synthesized from CT images in unpaired dataset can also be used to train the segmentor, as they are supposed to contain the same anatomical structure information as the input CT images. The corresponding loss function is defined as follows:

$$L_{\text{seg}}^{\text{s}} = \mathbb{E}_{(x,b)\sim(\mathcal{C}_u,\mathcal{B}_u)}\left[-b\,logS\left(G_{\text{C}\to\text{M}}(x)\right) - (1-b)\log\left(1-S\left(G_{\text{C}\to\text{M}}(x)\right)\right)\right] \tag{3.6}$$

where (x, b) denotes a CT patch and its corresponding segmentation annotation.

3.3.3 Neighbor-Based Anchoring Method

We utilize the cycle-consistent cross-modality image synthesis to transfer the annotation information from CT modality to MRI modality. However, recent study shows that the unpaired cycle-consistent image synthesis suffers from the ambiguity problem, which may lead to the geometric distortion [19]. Although we use both paired and unpaired MRI-CT data to train the image synthesis sub-network in a semi-supervised manner, the paired images have limited influence in the training since the number of paired images is typically much smaller than the unpaired images in clinical practice.

In this section, we introduce a novel neighbor-based anchoring method based on the paired MRI-CT data to alleviate the ambiguity problem of CycleGAN. The proposed method is based on an assumption that the *source and target domains (modalities) lie on respective manifolds with similar local geometry*. For example, if two patches in one modality are similar to each other in terms of semantic structure, their synthetic counterparts in another modality should also be similar to each other. Based on such assumption, we propose to reduce the space of feasible mappings in unsupervised cross-modality image synthesis by using the paired MRI-CT data as anchors.

Figure 3.3 illustrates the proposed neighbor-based anchoring method. As shown in the figure, two anchor point sets are built for the respective domains (i.e., CT and MRI domains), where the anchor points in two domains are actually the paired patches cropped from the paired MRI-CT images. Thus, for each patch (anchor point) in the CT anchor point set, there is a corresponding patch available in the MRI anchor point set and vice versa. Given an input patch x in the source domain (e.g., CT domain), we first find K nearest anchor points $\{x_1, x_2, \ldots, x_K\}$ for it in the same domain. Then the desired synthetic target y can be located near the anchor points $\{y_1, y_2, \ldots, y_K\}$ in the target domain (e.g., the MRI domain), where (x_i, y_i) $(i = 1, 2, \ldots, K)$ are paired anchor points in the source and target domains. In this way, we can effectively reduce the space of feasible mappings in unsupervised cross-modality image synthesis, thus alleviating the ambiguity problem of CycleGAN. The corresponding loss function is defined as follows:

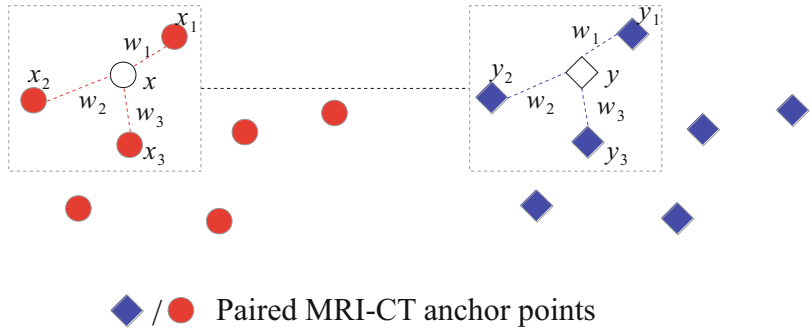


Fig. 3.3 The patches in two different domains (modalities) are assumed to lie on respective manifolds with locally similar geometry. Given an input x in the source domain (e.g., the left domain) and denote $\{x_1, x_2, x_3\}$ as the nearest anchor points of x, then desired synthetic target y in the target domain (e.g, the right domain) can be located by the $\{y_1, y_2, y_3\}$, where y_i $(i = 1, 2, 3)$ is the anchor point in the target domain corresponding to x_i

$$L_{\text{anchor}} = \mathbb{E}_{(x,y)\sim(\mathcal{C}_u,\mathcal{M}_u)} \frac{1}{K} \sum_{i=1}^{K} \left(w_i^{\text{C}} \parallel G_{\text{C}\to\text{M}}(x) - T(x_i) \parallel_1 + w_i^{\text{M}} \parallel G_{\text{M}\to\text{C}}(y) - T(y_i) \parallel_1 \right) \tag{3.7}$$

where $x \in \mathcal{C}_u$ and $y \in \mathcal{M}_u$, respectively, represent the unpaired CT and MRI patches and (x_i, y_i) denote the i-th neighboring anchor points of (x, y) in respective domains. The operator $T(\bullet)$ picks the corresponding anchor point in another domain for the given input (i.e., $T(x_i) = y_i$ and $T(y_i) = x_i$)). The weighting coefficients $w_i^{\text{C}} = \exp\left(-\alpha d_i^{\text{C}}\right)$ and $w_i^{\text{M}} = \exp\left(-\alpha d_i^{\text{M}}\right)$ are respectively used to balance the importance of each neighboring anchor point x_i and y_i for x and y based on the distances between them (i.e., $d_i^{\text{C}} = \parallel x - x_i^{\text{C}} \parallel_1$ and $d_i^{\text{M}} = \parallel y - y_i^{\text{M}} \parallel_1$), and α is a smoothing parameter.

The proposed neighbor-based anchoring method can effectively alleviate the ambiguity problem in unpaired cycle-consistent cross-modality image synthesis by making full use of the paired MRI-CT data. However, because paired data is typically very limited in clinical practice, it would impair the effectiveness of the proposed method since limitations in paired data cannot provide enough anchor points to cover the entire manifold. To address this issue, we propose to progressively extend the anchor point sets with synthetic patches from the unpaired dataset. More specifically, the anchor point sets are first initialized with the paired patches randomly cropped from the paired MRI-CT images. During the training stage, the distances between the unpaired patches and their nearest anchor points are recorded, and the top 20% patches that are farthest from their nearest anchor points are selected to augment the anchor point sets along with their synthetic results. In this way, we progressively extend the anchor point sets to expand the coverage area of anchor points on manifolds. Note that we only add the synthetic anchor points in the sparse area, i.e., far from the existing anchor points. These supplementary anchor points, although are not as good as those cropped from the real data, are still better than nothing in providing additional guidance for cross-modality image synthesis. The effectiveness of the proposed neighbor-based anchoring method will be further analyzed in Sect. 3.4.4 via ablation study.

3.3.4 Feature-Matching-Based Semantic Consistency

In our framework, the synthetic images are supposed to contain the same anatomical structure information as their corresponding inputs. However, this is not guaranteed in the original CycleGAN which only focuses on the pixel-level reconstruction at the visual image level and is insensitive to the anatomical structure information variations in high-level semantic space. To ameliorate

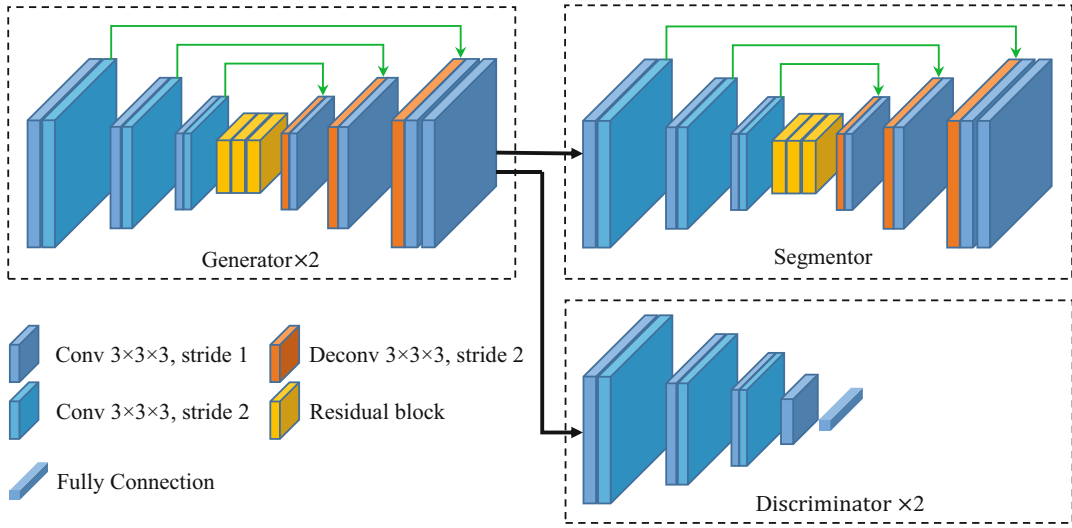


Fig. 3.4 Network architectures

this issue, we propose a feature-matching-based semantic consistent regularization to encourage the anatomical structure information preservation in synthesis. To be specific, we require the high-level anatomical structure information of the real and synthetic images to be consistent in synthesis. This is enabled by reducing the differences between feature maps extracted from the real and synthetic MRI images with the following loss function:

$$L_{\mathrm{sc}} = \mathbb{E}_{(x,y)\sim(\mathcal{C}_p,\mathcal{M}_p)} \| f_{\mathrm{S}}(y) - f_{\mathrm{S}}\left(G_{\mathrm{C}\to\mathrm{M}}(x)\right) \|_2^2 \tag{3.8}$$

where $x \in \mathcal{C}_p$ and $y \in \mathcal{M}_p$ denote a paired CT-MRI patches and $f_{\mathrm{S}}(\bullet)$ is an operator to extract the high-level semantic feature from the given input. In this work, such semantic feature is from the last but one activation layer (before the final convolutional layer) of segmentor S, which contains the critical anatomical structure information of the input data.

3.3.5 Network Architecture

The proposed framework consists of two generators, two discriminators and a segmentor. In this section, we introduce the detailed structure of each sub-network. For ease of understanding, the architecture of each sub-network is illustrated in Fig. 3.4.

3.3.5.1 Generators

The generators $G_{\mathrm{C}\to\mathrm{M}}$ and $G_{\mathrm{M}\to\mathrm{C}}$ learn the cycle-consistent mappings between the MRI and CT modalities. In this work, they are implemented to have the save architecture, i.e., a U-Net [20] like 3D fully convolutional neural network (FCN) which consists of (1) an encoder, (2) a decoder, (3) several long-range skip connections to link the encoder and decoder layers at the same resolution, level and (4) a series of residual blocks [21] to bridge the encoder and decoder. To be specific, the encoder comprises of three encoding blocks. Each encoding block consists of two $3 \times 3 \times 3$ convolutional layers with strides 1 and 2, respectively. The decoder is symmetrical to the encoder, which comprises of three decoding blocks and each block consists of a deconvolutional layer [22] and a convolutional layer. Similar to the 3D U-Net [20], three long-range shortcuts are applied to connect the encoder and decoder layers at every resolution levels to fasten the convergence as well as to achieve smoother results. Moreover,

three residual blocks [21] are used to bride the encoder and decoder to improve the nonlinear modeling capacity of the generator. The number of filters in generators starts at 32, doubles at each strided convolutional layer and halves at each deconvolutional layer. In addition, we apply the *Instance Normalization* [23] and *ReLU* activation [24] after each convolutional/deconvolutional layer except the last layer (the output layer), where we use the *tanh* activation to normalize the output.

3.3.5.2 Discriminators

The discriminators D_{M} and D_{C} distinguish the real images from the synthetic ones for MRI and CT modalities, respectively. They have a similar structure with the encoding part of generators, which also contain three encoding blocks. Besides, an additional convolutional layer and a fully connection layer are appended to generate the output with only one element, indicating whether the input image is real or synthetic. However, instead of the *ReLU* activation, we adopt the *Leaky ReLU* [25] activation with slope 0.2 for discriminators as suggested in DCGAN [7].

3.3.5.3 Segmentor

Our segmentor is very similar to the generator, as shown in Fig. 3.4. However, we use the *softmax* activation instead of the *tanh* activation after the output layer of segmentor to predict the probability of each voxel belonging to each class (bone or background).

3.3.6 Implementation Details

Due to GPU memory limitation, we used small 3D patches with the size of $32 \times 32 \times 32$ to train our model, where the patches were randomly cropped from MRI and CT images. Both paired and unpaired data were used to train the model with a batch size of four, where each mini-batch consists of two paired and two unpaired MRI-CT patches. We split all sub-networks into three groups, i.e., the generators' group, the discriminators' group, and the segmentor's group, and trained them alternately in an adversarial manner at each iteration. To be specific, the discriminators D_{M} and D_{C} were first trained using loss function (3). Then the generators $G_{\mathrm{C}\to\mathrm{M}}$, $G_{\mathrm{M}\to\mathrm{C}}$ were trained by minimizing the following loss function:

$$\mathrm{L}_{\mathrm{pair}} + \lambda_1 \mathrm{L}_{\mathrm{unpair}} + \lambda_2 \mathrm{L}_{\mathrm{adv}} + \lambda_3 \mathrm{L}_{\mathrm{anchor}} + \lambda_4 \mathrm{L}_{\mathrm{sc}} \tag{3.9}$$

Finally, the segmentor was trained by using both real and synthetic data as follows:

$$L^{\mathrm{r}}_{\mathrm{seg}} + \lambda_5 L^{\mathrm{s}}_{\mathrm{seg}} \tag{3.10}$$

We trained all sub-networks with the above-mentioned strategy for a total of 100,000 iterations. The *Adam* [26] was adopted as the optimizer to train the model with a learning rate of 10^{-4}. Other hyper-parameters were empirically set as $K = 3, \alpha = 1, \lambda_1 = 0.5, \lambda_2 = 0.05, \lambda_3 = 0.1, \lambda_4 = 0.5$, and $\lambda_5 = 0.5$.

3.4 Experiment and Results

To demonstrate the effectiveness of the proposed method, we conducted the experiments on CMF bony structure segmentation and compared the proposed method with the state-of-the-art medical image segmentation methods. In this section, we first elaborate the experimental settings in Sect. 3.4.1, including the training and testing datasets, the preprocessing procedure, and the evaluation principle. Then we visually present some representative cross-modality image synthesis results of the proposed method in Sect. 3.4.2. The segmentation performance of the proposed method is evaluated both qualitatively and quantitatively in Sect. 3.4.3 and is compared with the state-of-the-art medical image segmentation methods to demonstrate the superiority of the proposed method. Finally, we make an in-depth study of the proposed method by quantitatively analyzing its two key components (i.e., the neighbor-based anchoring method and the feature-matching-based semantic consistency constraint) via ablation study in Sect. 3.4.4.

3.4.1 Experimental Setting

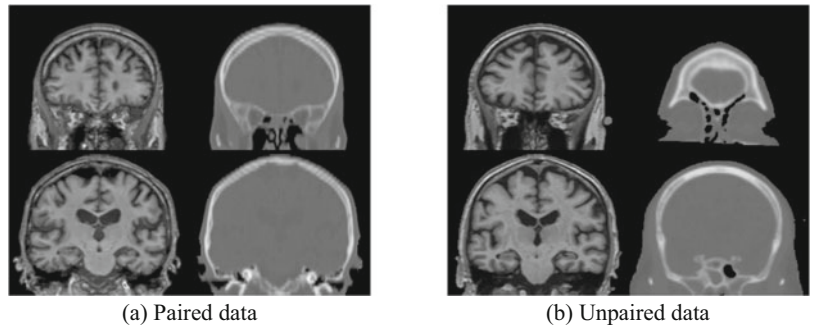
We used both paired and unpaired MRI-CT datasets in our experiments. The paired dataset consists of eight pairs of MRI-CT scans from the Alzheimer's disease Neuroimaging Initiative (ADNI) dataset [27]. The unpaired dataset contains 50 MRI scans and 50 CT scans, which were randomly selected from the ADNI dataset and the CQ500 dataset [28], respectively. All images were aligned to a common template using the FMRIB's Linear Image Registration Tool (FLIRT) and cropped to have the same size of $146 \times 180 \times 175$ with a voxel size of $1 \times 1 \times 1$ mm ^{3}. Then, the intensity values in all images were linearly scaled into the range of $[-1, 1]$. Examples of paired and unpaired MRI-CT images are shown in Fig. 3.5. Since the bony boundaries are clearly shown in CT images, the ground truth of bones was achieved by thresholding and then manually refined by an expert as needed.

We evaluated the performance of the proposed method in the scenario where the annotated data was very limited. To be specific, the proposed model was trained and validated eight times, corresponding to the number of paired MRI-CT images. At each time, only a single MRI-CT pair from the paired dataset was selected to train the model along with all unpaired data, leaving the other seven pairs for testing.

3.4.2 Image Synthesis Results

We first qualitatively evaluated the cross-modality image synthesis performance of the proposed method by presenting some representative visual results of synthetic images. For comparison purposes, we trained the image synthesis sub-network (i.e., two generators and two discriminators) separately using only paired data and refer to it as the *Baseline* method. Moreover, we also compared the proposed method with the well-known CycleGAN method which was trained by using only unpaired data. The comparison results are shown in Fig. 3.6, where five columns respectively represent (a) the inputs, (b) the results of the Baseline method, (c) the results of the CycleGAN method, and (d) the results of the proposed method and the ground truth. Compared with the baseline, the proposed method was able to produce more clear and realistic results in all testing data. This can be attributed to the proposed semi-supervised framework which learns data distribution from both paired and unpaired images. On the other hand, the CycleGAN yielded much poorer results, especially for soft tissues. This is reasonable since the 3D small (unpaired) MRI-CT patches can hardly provide enough contextual information to learn reliable cross-modality mappings, even if the cycle-consistent regularization was applied.

In addition, from the results in Fig. 3.6, we can also observe that learning the CT-to-MRI translation is much more challenging than learning the MRI-to-CT translation, especially for soft tissues. In fact, it is very difficult (if at all possible) to reliably reconstruct the soft tissues from CT images since in which the soft tissues are almost invisible. However, it is worth noting that the CMF surgery is a bony surgical procedure, so we only care about the bony areas, and the soft-



(a) Paired data (b) Unpaired data

Fig. 3.5 Examples of paired and unpaired data

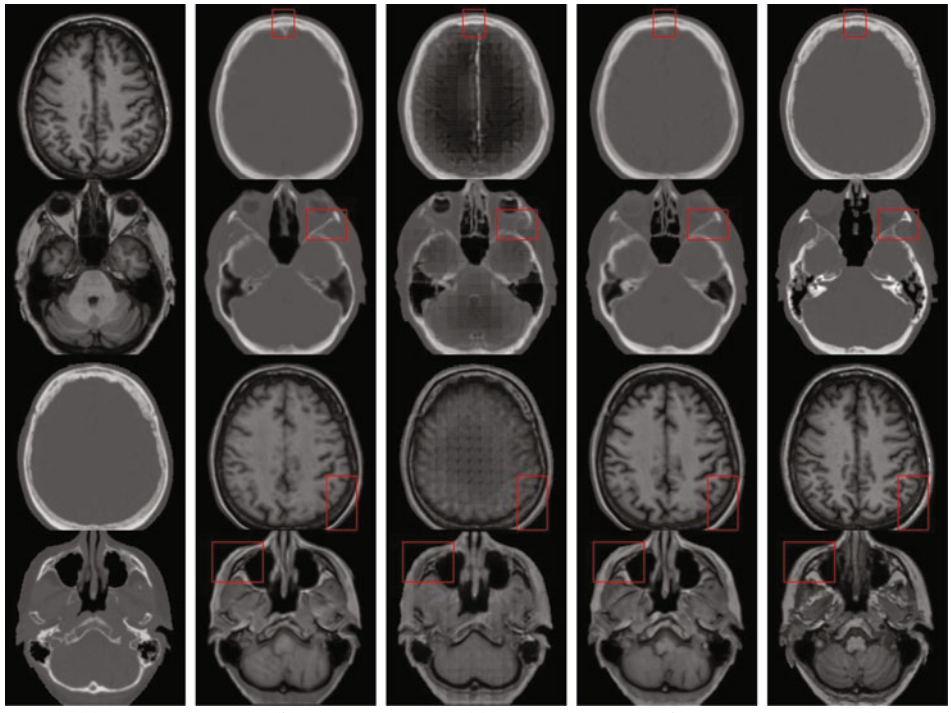


Fig. 3.6 Visual results of cross-modality image synthesis

tissue information is unnecessary. The segmentation performance of the proposed method will be evaluated in the following section.

3.4.3 Segmentation Performance

We also evaluated the segmentation performance of the proposed method. To demonstrate the superiority of the proposed method, we compared it with two state-of-the-art medical image segmentation methods, i.e., the well-known *3D U-Net* [20] and the recent proposed *Deep-supGAN* [2], which were exclusively trained on paired dataset. Furthermore, we implemented two variants of the proposed method for comparison: (1) the *Baseline* method which trained the segmentor solely with only paired data and (2) the *Two-Stage* method that trained the image synthesis sub-network and the segmentor independently in sequence instead of training them jointly in an adversarial manner. The segmentation performance of all comparison methods was quantified in terms of Dice Similarity Coefficients (DSC) and Average Symmetric Surface Distance (ASSD).

The comparison results are shown in Table 3.1, where the bold numbers indicate the best results. From Table 3.1, it is obvious that the Baseline, 3D U-Net, and Deep-supGAN methods yielded much poorer results, implying that they suffer from the overfitting problem since the training data (i.e., the paired MRI-CT data) is very limited. In contrast, the proposed method effectively made use of the synthetic data to augment the training dataset via cross-modality image synthesis, leading to the performance improvement in terms of both DSC (85.66 ± 3.26) and ASSD (1.04 ± 0.19), respectively. Additionally, the proposed method also outperforms its Two-Stage variant model by 1.81 and 0.08 in terms of DSC and ASSD, respectively. This can be attributed to the proposed training strategy, which trained the image synthesis sub-network and the segmentation sub-network jointly in an adversarial manner.

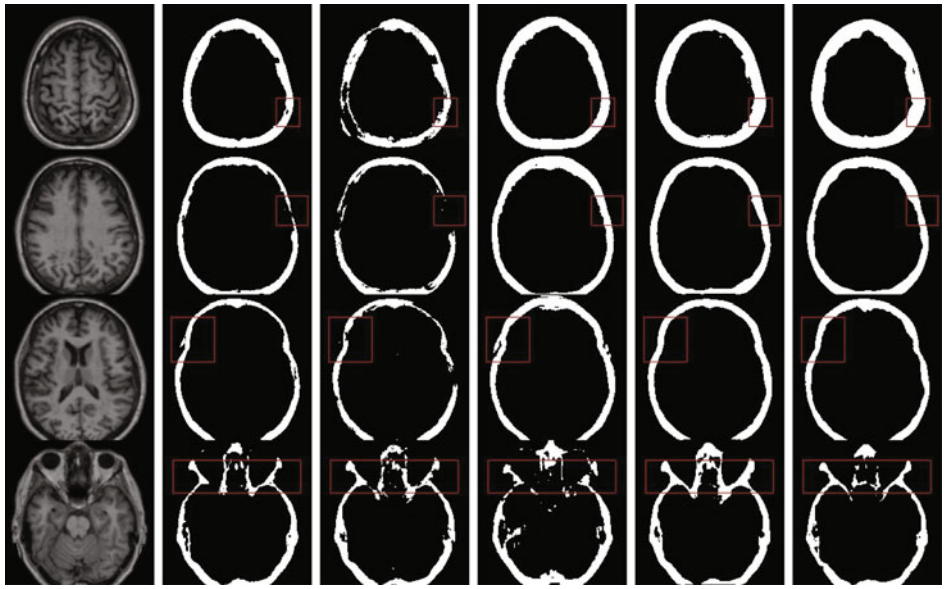


Fig. 3.7 Visual results of bony structures segmentation

We also qualitatively compared the segmentation results of different methods. The visual segmentation results are shown in Fig. 3.7, where the six columns respectively represent (a) the input MRI images, (b) the results of the Baseline method, (c) the results of the 3D U-Net method, (d) the results of the Deep-supGAN method, (e) the results of the proposed method, and (f) the ground truth. It is obvious that the proposed method achieved the best results. Compared with other competing methods, the bony structures predicted by the proposed method are more complete and smoother, as indicated by the red boxes. This is consistent with the quantitative results presented in Table 3.1, demonstrating the superiority of the proposed method.

Table 3.1 Comparison of segmentation performance

Method	DSC (%)	ASSD
Baseline	78.68 ± 5.42	1.33 ± 0.25
3D U-Net [20]	77.41 ± 5.41	1.34 ± 0.20
Deep-supGAN [2]	77.82 ± 3.18	1.49 ± 0.20
Two-Stage	83.85 ± 4.03	1.12 ± 0.22
Proposed	**85.66 ± 3.26**	**1.04 ± 0.19**

Mean ± standard deviation

3.4.4 Ablation Study

To make an in-depth study of the proposed method, we conducted the ablation study to quantitatively analyze two key components of the proposed method, i.e., the neighbor-based anchoring method and the feature-matching-based semantic consistency constraint. To be specific, two variants of the proposed method were trained without using the neighbor-based anchor method (denoted as *w/o Anchor*) and the feature-matching-based semantic consistency constraint (denoted as *w/o SC*), respectively. The segmentation performance of these variant models is evaluated and compared with the original method under the same evaluation principle. From the comparison results in Table 3.2, we can observe that the proposed anchoring method and the segmentation consistency constraint successfully alleviate the ambiguity problem and encourage the semantic consistency in cross-modality image synthesis, leading to the performance improvement.

Table 3.2 Ablation study

Methods	DSC (%)	ASSD
W/o anchor	84.81 ± 3.44	1.08 ± 0.20
W/o SC	84.27 ± 3.54	1.16 ± 0.20
Proposed	**85.66 ± 3.26**	**1.04 ± 0.19**

Mean ± standard deviation

3.5 Conclusion and Discussion

In this work, we introduced a method to automatically segment CMF bony structures from MRI. To address the annotation (i.e., label) limitation problem, we proposed to transfer the annotation information from CT modality to MRI modality via cycle-consistent cross-modality image synthesis. The synthetic MRI images were supposed to contain the same anatomical structure information as their corresponding inputs (i.e., CT images) and thus could be used to augment the training dataset for segmentation. To alleviate the ambiguity problem, we proposed a novel neighbor-based anchoring method to reduce the space of possible mappings in cross-modality image synthesis by making use of the limited paired data. Moreover, a feature-matching-based semantic consistency constraint was proposed to encourage a segmentation-oriented image synthesis by enforcing the high-level semantic consistency between the real and synthetic images. The qualitative and quantitative experimental results demonstrated the superiority of the proposed method in comparison with the state-of-the-art medical image segmentation methods. We also conducted the ablation study to further analyze the effects of the proposed neighbor-based anchoring method and the feature-matching-based semantic consistency constraint.

References

1. Nie D, Wang L, Trullo R, Li J, Yuan P, Xia J, Shen D. Segmentation of craniomaxillofacial bony structures from mri with a 3d deep-learning based cascade framework. In: International workshop on machine learning in medical imaging. Cham: Springer; 2017. p. 266–73.
2. Zhao M, Wang L, Chen J, Nie D, Cong Y, Ahmad S, Ho A, Yuan P, Fung SH, Deng HH, Xia J. Craniomaxillofacial bony structures segmentation from MRI with deep-supervision adversarial learning. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2018. p. 720–7.
3. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision; 2017. pp. 2223–2232.
4. Kim T, Cha M, Kim H, Lee JK, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70; 2017 Aug 6. pp. 1857–1865. JMLR. org.
5. Al Hasan M, Chaoji V, Salem S, Zaki M. Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security; 2006 Apr 20.
6. Møller MF. A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. 1993 Jan 1;6(4):525–33.
7. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434. 2015 Nov 19.
8. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach Learn. 2001 Jan 1;42(1–2):177–96.
9. Bennett KP, Demiriz A. Semi-supervised support vector machines. In Advances in Neural Information processing systems; 1999. pp. 368–374.
10. Weston J, Ratle F, Mobahi H, Collobert R. Deep learning via semi-supervised embedding. In: Neural networks: tricks of the trade. Berlin, Heidelberg: Springer; 2012. p. 639–55.
11. Bai W, Oktay O, Sinclair M, Suzuki H, Rajchl M, Tarroni G, Glocker B, King A, Matthews PM, Rueckert D. Semi-supervised learning for network-based cardiac MR image segmentation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2017 Sep 10. p. 253–60.
12. Mahapatra D. Semi-supervised learning and graph cuts for consensus based medical image segmentation. Pattern Recogn. 2017 Mar 1;63:700–9.
13. Huang Y, Beltrachini L, Shao L, Frangi AF. Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging. In: International workshop on simulation and synthesis in medical imaging. Cham: Springer; 2016 Oct 21. p. 118–26.
14. Van Nguyen H, Zhou K, Vemulapalli R. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2015 Oct 5. p. 677–84.

15. Hiasa Y, Otake Y, Takao M, Matsuoka T, Takashima K, Carass A, Prince JL, Sugano N, Sato Y. Cross-modality image synthesis from unpaired data using CycleGAN. In: International workshop on simulation and synthesis in medical imaging. Cham: Springer; 2018 Sep 16. p. 31–41.
16. Huang Y, Shao L, Frangi AF. Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. pp. 6070–6079.
17. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision; 2017. pp. 2794–2802.
18. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In Advances in neural information processing systems; 2014. pp. 2672–2680.
19. Zhang Z, Yang L, Zheng Y. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. pp. 9242–9251.
20. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Cham: Springer; 2016 Oct 17. p. 424–32.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. pp. 770–778.
22. Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. In ICCV; 2011. (Vol. 1, No. 2, p. 6).
23. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022. 2016 Jul 27.
24. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10); 2010. pp. 807–814.
25. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. In Proc. icml; 2013 Jun 16 (Vol. 30, No. 1, p. 3).
26. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.
27. Trzepacz PT, Yu P, Sun J, Schuh K, Case M, Witte MM, Hochstetler H, Hake A. Alzheimer's disease neuroimaging initiative. Comparison of neuroimaging modalities for the prediction of conversion from mild cognitive impairment to Alzheimer's dementia. Neurobiol Aging. 2014 Jan 1;35(1):143–51.
28. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018 Dec 1;392(10162):2388–96.

4 Sparse Dictionary Learning for 3D Craniomaxillofacial Skeleton Estimation Based on 2D Face Photographs

Deqiang Xiao, Chunfeng Lian, Li Wang, Hannah Deng, Kim-Han Thung, Pew-Thian Yap, James J. Xia, and Dinggang Shen

4.1 Introduction

Craniomaxillofacial (CMF) defects can be caused by traffic accidents, combat injuries, tumor ablations, etc. CMF reconstructive surgery is specifically designed to reconstruct such defects, by restoring the defected facial bones to a normal shape or a desired pre-traumatic status. The success of the CMF reconstructive surgery largely depends on accurate surgical planning [1].

In the conventional clinical practices of CMF surgical planning, computed tomography (CT) or cone beam computed tomography (CBCT) scans of the patient's head are always acquired before surgical operations, so that a three-dimensional (3D) bone model can be extracted from the acquired (CB)CT scans. After that, a surgeon can simulate the surgery by virtually cutting the 3D model into multiple bony segments and moving the segments to their desired positions. The desired positions are computed by comparing the values of the patient's cephalometric measurements with the corresponding normal values calculated on a set of normal subjects. The cephalometric measurements consist of a group of distances and angles computed from several anatomical landmarks localized on the facial bone. However, as the normal values used in the above-mentioned procedure are a set of average values, they cannot provide accurate guidance for a specific patient. Furthermore, as the conventional cephalometric measurements are conducted in two-dimensional (2D) space, they might result in large errors when they are used to simulate a 3D model.

Since 2D cephalometric analysis is limited for providing accurate guidance on CMF surgical planning, a 3D computer-aided surgical simulation (CASS) system for different CMF surgeries was proposed by Xia et al. [2]. In CASS, the patient's facial bone model is first oriented in

D. Xiao (✉) · C. Lian · L. Wang · K.-H. Thung · P.-T. Yap · D. Shen
BRIC and Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, USA

H. Deng · J. J. Xia
Department of Oral and Maxillofacial Surgery, Houston Methodist Hospital, Houston, TX, USA

J. J. Xia
Department of Surgery (Oral and Maxillofacial Surgery), Weill Medical College, Cornell University, NY, Ithaca, USA
e-mail: jxia@houstonmethodist.org

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_4

the natural head position (NHP), followed by a 3D cephalometric analysis to estimate the CMF deformities [3]. As 3D information is used during the cephalometric analysis, the performance of the surgical planning using CASS is improved when compared to the conventional 2D cephalometric analysis-based methods.

Although cephalometric analysis-based planning works well in terms of both accuracy and cost efficiency in treating regular jaw deformities [4], it is still challenging to produce ideal surgical outcomes for patients with complex CMF defects. One potential reason is that the current CASS quantifies deformities subjectively, as it heavily depends on the surgeon's experience. Nonetheless, it is technically impossible to quantify and restore patient-specific defects (e.g., trauma) referring only to the simple mean numerical values of limited cephalometric evaluations on normal subjects.

Intuitively, estimating patient-specific normal facial bone can provide a paradigm-shift avenue in reconstructive surgery for patients with CMF defects. One of the most commonly used methods for patient-specific facial bone estimation is mirror imaging mapping [5, 6], typically realized by mapping the facial skeleton from the normal side to the defected side. Since the symmetric mapping is based on the too-strict hypothesis of symmetric human facial structure, it is very limited and cannot handle the cases of bilateral defects. Statistical shape model-based (SSM-based) normal facial bone estimation is another commonly used method [7–9]. In SSM-based methods, a set of normal facial bone shapes are first acquired from healthy subjects, upon which the principal component analysis (PCA) [10] is applied to construct an SSM [11]. Then, the patient's normal bone shape is estimated by fitting the established SSM onto the remaining normal parts of the patient's facial bone. A key limitation of the SSM-based method is its weak generalization capacity due to the practical challenge that the dataset available for establishing SSM is usually small. Recently, methods using geometric deformation were also proposed to estimate normal facial bone [12–14]. They estimate normal bone shape by deforming the patient's defected bone with the deformation field calculated by the surface interpolating techniques of thin plate spline (TPS) [15] or Laplacian surface editing [16]. Generally, these geometric deformation-based methods can produce accurate normal bone estimations. However, they cannot handle large-scale defects, since the deformation fields quantified by using limited anatomical landmarks and surface interpolating techniques are relatively coarse.

In [17], Xiao et al. assumed that a personalized reference bone shape can be estimated from pre-traumatic face photographs of the patient with large-scale CMF defects. That is, the patient's pre-traumatic 2D portrait photographs can be used to estimate a 3D reference bone model, in which the normal facial bones are retained, and the defected facial bones are restored. To this end, a set of head (CB)CT images are acquired from normal subjects to extract both their facial and bony surfaces. They form a dictionary (or a set of atlases), for which the dense correspondences are established across subjects through nonrigid surface matching, and a correlation model is constructed between facial and bony surfaces. For a given patient, a set of pre-traumatic 2D face photographs (i.e., common daily life photos) are first collected, based on which a 3D face surface is reconstructed. Afterwards, by feeding the reconstructed 3D face (from 2D photographs) into the constructed correlation model, a normal bone shape model is generated for the patient via spare dictionary learning [18]. However, the generated shape model is a relatively coarse (initial) estimation, since the 3D face reconstructed from previous photographs may not necessarily reflect the patient's current face shape and might lead to an inaccurate output of the correlation model. To further improve the accuracy of the reference shape estimation, the patient's current facial bone (posttraumatic bone) is used to refine the initial estimation. The refinement is achieved by deforming the initially estimated normal bone surface onto the patient's posttraumatic bone surface with a deformable shape model. Additionally, to guarantee the normality of the estimated facial bone, a statistical bone shape model constructed from the normal subjects is applied to constrain the deformable shape model applied on the estimated facial bone to avoid overfitting. The final

reference shape model after refinement can then be applied to guide the CMF surgical planning.

In this chapter, we will introduce this machine learning-based reference model estimation method in detail. The rest of the chapter is organized as follows. We introduce details of this method in Section 2 and give corresponding experiments with results in Section 3. Also, a summary is made in Section 4.

4.2 Method

The proposed method, summarized in Fig. 4.1, consists of three major steps: (1) reconstruction of the pre-traumatic 3D facial surface, (2) estimation of the initial reference bony shape model, and (3) refinement of the bony shape model.

4.2.1 Patient's Normal 3D Face Reconstruction

The workflow for normal 3D face reconstruction from a set of the patient's pre-traumatic 2D face photographs is shown in Fig. 4.2. Since a 2D face photograph provides limited information for accurate 3D face reconstruction, a set of 3D face key points for each 2D face photograph is first generated. To achieve this, the CNN-based algorithm [20] is utilized to automatically extract 68 3D face key points from a face photograph. Then, based on the detected 3D face key points, the patient's normal face surface is derived from the mean face in Basel face model (BFM) [21]. Specifically, for each set of 68 3D face key points reconstructed from a single face photograph, a sparse deformation field is generated between the reconstructed key points and the corresponding face key points on the BFM mean face. The calculated sparse deformation field was then converted into a dense deformation field by thin plate spline (TPS) interpolation [15]. By applying the dense deformation field on the BFM mean face, the patient's normal face surface is estimated. As each 2D face photo can be used to generate a normal face surface, the face reconstruction accuracy is further improved by applying the method proposed in [22] to merge all reconstructed face surfaces into a single face surface.

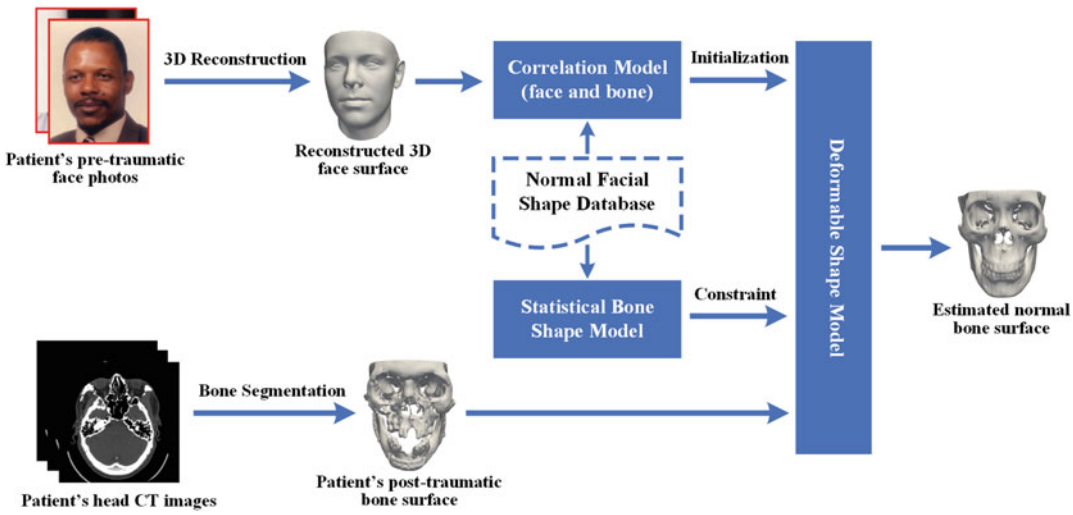


Fig. 4.1 The overview workflow of our proposed method. The method mainly contains three steps: First, a 3D facial surface is reconstructed from the patient's pre-traumatic 2D face photos. Second, we construct a correlation model between face and bone from a normal facial shape database and generate an initial estimation from the correlation model under the input of reconstructed 3D face (from 2D photos). Third, by deforming the initially estimated normal bone onto the patient's current bone (from posttraumatic CT images), we finally get the refined normal estimation. The figure is from Xiao et al. [19]

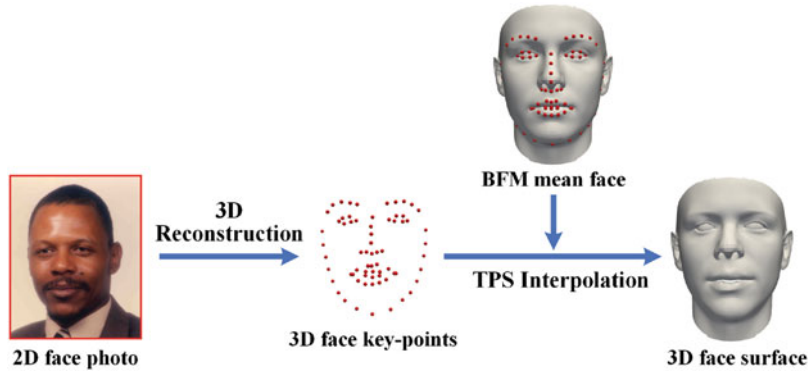


Fig. 4.2 The illustration of 3D face reconstruction from 2D face photo. The figure is from Xiao et al. [19]

4.2.2 Sparse Dictionary Learning for Initial Reference Bone Model Estimation

Given a reconstructed face surface (as described in Section 2.1), an initial estimate of the corresponding bone surface can be generated based on a correlation model between the face and bone surfaces, under the assumption that these two surfaces are highly correlated. Specifically, a normal facial shape database is first constructed. The database consists of face and bone surfaces extracted from a set of head CT scans of normal subjects (see Fig. 4.3). Then, the sparse representation technique [18] is used to estimate the reconstructed face surface of a patient with the normal face surfaces in the database. Finally, by applying the learned sparse coefficients on the corresponding normal bone surfaces, the patient's normal bone structure is estimated. The detailed steps in obtaining the initial bone surface estimate are described as follows.

Normal Facial Shape Database Construction The normal facial shape database consists of face and bone surfaces extracted from a set of CT scans of normal subjects. Specifically, using a segmentation algorithm [23] on (CB)CT images, the facial soft tissue and bone structure are extracted from the head CT scans for each normal subject. Then, based on the segmented results, both the face and bone 3D surface models are generated by using the marching cubes algorithm [24]. The generated (face and bone) surfaces from different subjects are then properly aligned for the subsequent correlation model. That is, a facial landmark detection method [25] is first used to extract landmarks on each facial bone structure, and then these landmarks are applied to rigidly align all extracted shapes (i.e., face and bone

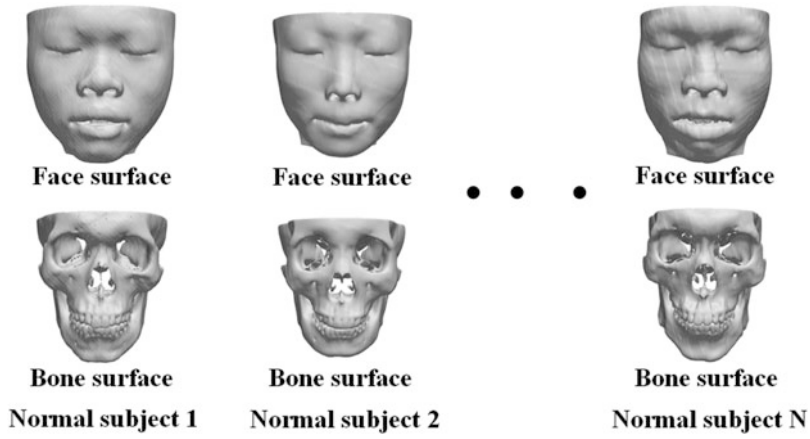


Fig. 4.3 The normal facial shape database, consisting of face and bone surfaces extracted from a collection of normal subjects' CT images. The figure is from Xiao et al. [19]

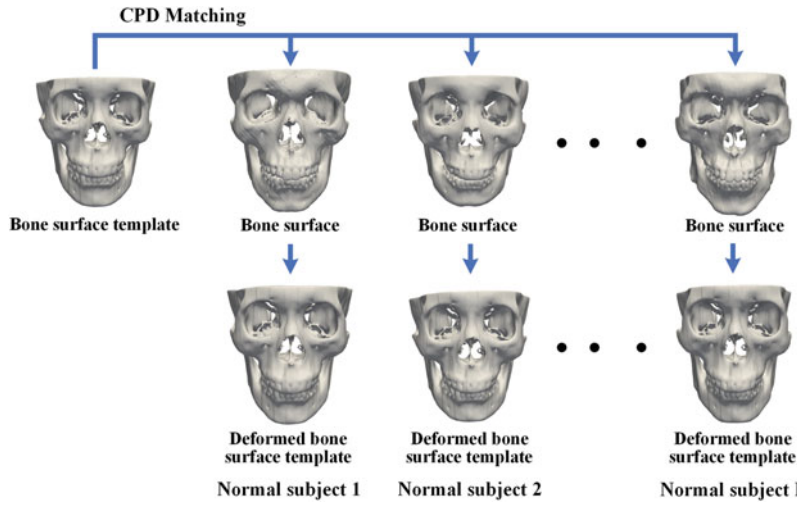


Fig. 4.4 Searching corresponding points on bone surfaces by deforming the bone surface template onto the rest normal bone surfaces with the CPD matching algorithm. The figure is from Xiao et al. [19]

surfaces) from different subjects together (see Fig. 4.3). Furthermore, to establish dense correspondences for the surface points on the extracted shapes across different subjects, nonrigid mapping from a template shape onto each aligned shape is performed by using the coherent point drift (CPD) algorithm [26] (see Fig. 4.4). The template shape is defined as the aligned shape that is closest to the mean shape of all the aligned surfaces. Once the dense correspondences between the template and aligned shapes are established, two dictionaries (i.e., one for the face surface and one for the bone surface) containing all the corresponding surface points are built.

Sparse Dictionary Learning The aim of sparse dictionary learning is to search a representation (sparse coding) of the input data in the form of linear combination of atoms in a dictionary, in which the representation is computed as a set of sparse coefficients (the majority of coefficients are zeros) for the corresponding atoms in the dictionary. Thus, dictionary construction and sparse coding are two main tasks in this procedure. Usually, we collect the training data to represent each training sample as a vector and stack all vectors into a matrix as the dictionary. The sparse coding is realized by minimizing a cost function between input data and its representation from the dictionary to find an optimized set of sparse coefficients. The sparse dictionary learning has been successfully applied in the fields of image denosing, classification, compressing, etc. In the following paragraphs, we will introduce how to apply this machine learning technique into CMF skeleton estimation.

Correlation Model Construction Based on the hypothesis of linear correlation between the face and bone surfaces, the sparse representation technique [18] is applied to establish a correlation model between them. To this end, based on the surface corresponding points (between the template and aligned shapes) for all normal subjects in the facial shape database, two dictionaries (i.e., the face and bone dictionaries) are first constructed and then utilized in the correlation model. Specifically, the coordinates of all the corresponding points on the aligned face surfaces are stacked into a $3M \times N$ matrix as the face dictionary D_{Face}, where M is the number of corresponding points on each face surface and N is the number of normal subjects. Each column of the matrix corresponds to the (vectorized) coordinates of M corresponding points of one subject from the database. As three values are used to define the coordinate of a point in the 3D space, the number of elements in each column is $3M$. Similarly, the coordinates of all the corresponding points on the aligned bone surfaces in the database are stacked together into a matrix as the bone dictionary D_{Bone}. Then, given a coor-

dinate vector P_{Est}^{F} of the corresponding points on a patient's estimated normal face surface, a sparse representation model is used to sparsely represent it with the column vectors in the face dictionary D_{Face}. The sparse coefficient vector C is calculated by solving the following objective:

$$C_{min} = \arg\min_{C} \left\| D_{Face}C - P_{Est}^{F} \right\|_2^2 + \lambda_1 \|C\|_1 + \lambda_2 \|C\|_2, \quad (4.1)$$

where the first term is the data fitting term, $\|\bullet\|_2$ denotes the $l2$ norm, $\|\bullet\|_1$ denotes the $l1$ norm, and λ_1 and λ_2 are the two regularization parameters used to control the sparsity of C. Equation (4.1) is similar to the well-studied elastic net problem [27], which can be solved using the method proposed in [27]. With the calculated sparse coefficient vector C_{min}, the patient's normal bone surface points P_{Est}^{B} are estimated by

$$P_{Est}^{B} = D_{Bone}C_{min}. \quad (4.2)$$

under the assumption that the face and respective bone surfaces are highly correlated. Then, the patient's normal bone surface model can be derived from the bone template surface mesh based on the estimated points.

Initial Normal Bony Shape Estimation According to the established correlation model, the correspondence between the normal face estimated from 2D photographs (see Section 2.1) and the CT face template is searched. As the estimated normal face and the CT face template are acquired from different imaging spaces, mapping the estimated face onto the CT face template image space is performed using the iterative closest point (ICP) algorithm [28] with a similarity transform (rigid and scale transform). Then, nonrigid surface matching between the mapped face and the CT face template is realized using the CPD algorithm, where, for each sampled point on the CT face template, the corresponding point on the mapped face is obtained. Finally, by inputting the coordinate vector of the corresponding points (i.e., P_{Est}^{F}) into the established correlation model (i.e., Eq. (4.1) and (4.2)), an initial normal bone shape is obtained for the patient. This procedure is illustrated in Fig. 4.5.

4.2.3 Refinement of the Initially Estimated Reference Model

To refine the initially estimated reference bone shape, the adaptive-focus deformable shape model (AFDSM) [29], a variant of the snake model [30], is utilized to deform the initial estimation onto the patient's posttraumatic facial bone. The AFDSM-based nonrigid surface matching is realized by first defining an attribute vector on each vertex, so that the vertex

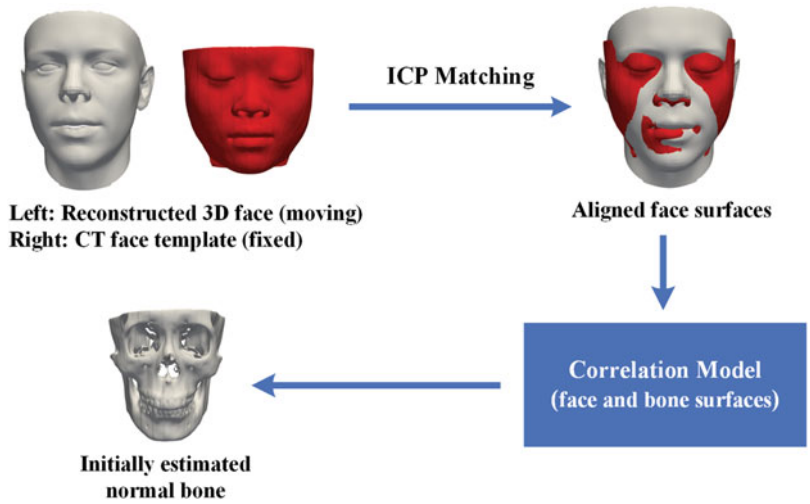


Fig. 4.5 The illustration of the workflow for generating the initial normal bone estimate from a reconstructed face surface. The figure is from Xiao et al. [19]

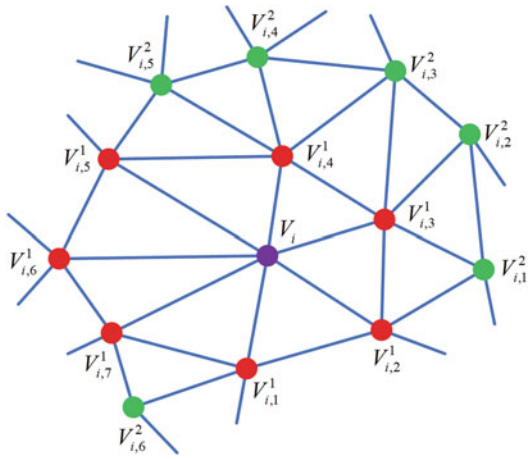


Fig. 4.6 Illustration of organized neighboring vertices of the vertex V_i. The red dots denote the neighboring vertices in the first layer, and the green dots denote the neighboring vertices in the second layer. The figure is from Xiao et al. [19]

differences before and after the deformation could be computed. To construct the attribute vector, the neighboring vertices of each vertex V_i are first organized into different layers on the surface mesh, where each neighboring vertex $V^k_{i,j}$ in the k^{th} layer is connected to V_i by k edges. Figure 4.6 illustrates the organized neighbors of a vertex on a surface mesh. Then, the attribute vector F_i of vertex V_i is defined as

$$\begin{aligned} F_i &= \frac{\left[f_{i,1}\ f_{i,2}\ f_{i,3}\cdots f_{i,R}\right]^T}{\sum_{i=1}^{N}\sum_{k=1}^{R}\left|f_{i,k}\right|}, \\ f_{i,k} &= \begin{vmatrix} x_i & x^k_{i,1} & x^k_{i,2} & x^k_{i,3} \\ y_i & y^k_{i,1} & y^k_{i,2} & y^k_{i,3} \\ z_i & z^k_{i,1} & z^k_{i,2} & z^k_{i,3} \\ 1 & 1 & 1 & 1 \end{vmatrix}, \end{aligned} \tag{4.3}$$

where $f_{i,k}$ denotes the determinant of a 4×4 matrix which contains position information of vertex V_i (i.e., $[x_i\ y_i\ z_i]$) and its three nearest neighbors (i.e., $\left\{\left[x^k_{i,j}\ y^k_{i,j}\ z^k_{i,j}\right], j = 1, 2, 3\right\}$) within the k^{th} layer, R denotes the number of neighboring layers for V_i, and N is the total number of vertices on the surface mesh. Generally, F_i reflects the local shape information when R is small and gradually captures more global shape information as R increases. Based on the attribute vector F_i, the energy function is defined as

$$E = \sum\nolimits_{i=1}^{N}\left(E_i^{model} + E_i^{data}\right), \tag{4.4}$$

where E_i^{model} denotes the degree of attribute vector difference between the initial and its deformed version for vertex V_i and E_i^{data} denotes the degree of attribute vector difference between the deformed and target shapes for vertex V_i. The initial and target shapes are defined as the initially estimated bone surface and the patient's posttraumatic bone surface, respectively. In other words, by minimizing energy in Eq. (4.4), the refined normal bone surface did not deviate too much from both the initial estimated and patient's posttraumatic bone surfaces. Furthermore, a greedy deformation algorithm [31] is applied to minimizing the energy function E in Eq. (4.4) based on affine transform.

To guarantee the normality of the deformed shape after each optimization iteration, the deformation procedure is constrained to be within a statistical normal shape. Specifically, a statistical shape model (SSM) [11] of bone shape is constructed from the normal facial shape database (see Section II.B) via PCA [10], such as

$$S^i_{SSM} = \overline{S} + W^i P, \tag{4.5}$$

where S^i_{SSM} denotes the reconstructed bone shape by applying the SSM on S^i, S^i denotes the deformed bone shape at the i^{th} iteration, $\overline{S}$ denotes the mean aligned bone shape in the normal facial bone shape database, W^i is a coefficients vector, and P is a matrix of principal components of normal bone shape. Here, each shape is represented as a vector of coordinates of all vertices on a bone surface model. Then, the constructed SSM is used to constrain each iteration of the optimization to avoid deformation overfitting, such as

$$S^i_{updated} = \alpha_i S^i_{SSM} + (1 - \alpha_i) S^i, \tag{4.6}$$

where $S^i_{updated}$ is the correspondingly updated bone shape and α_i is a hyperparameter between

zero and one that determines the weight between S^i_{SSM} and S^i in the i^{th} iteration. By gradually reducing α_i at each iteration, the last deformed shape is closer to the target shape. At each iteration, the deformed shape is updated and then used as input for the next iteration.

4.3 Experiments and Results

4.3.1 Experimental Dataset

A set of CT scans of 30 normal subjects are collected to construct the normal facial shape database. For image preprocessing, the images are segmented to extract face and bone surfaces. Fifty-one landmarks on facial bone also are detected for each subject.

The experimental data are split into two parts: synthetic patient data and real patient data. The synthetic patient data are generated from CT images of normal subjects. To simulate pre-traumatic 2D face photographs, the BFM mean face surface is deformed onto a normal face CT surface using the CPD algorithm, and then a statistical face texture model [21] is applied on the deformed face surface to set a color value on each surface vertex. Afterwards, the deformed face surface is rendered with the skin texture in 3D space for visualization. We adjust different poses of the displayed 3D face and take its 2D screen shots to get a group of face photographs (see Fig. 4.7). The synthetic posttraumatic bone shape is generated by manually modifying a normal CT bone structure. Specifically, an experienced CMF surgeon is recruited to remove and deform bone segments on a normal CT bone structure to simulate acquired CMF defects. For the real patient data, the CT scans from patients who suffered from severe CMF defects are collected, and the patients' two or three face photographs also are acquired. In addition, image segmentation and bone landmark digitization are performed on real patient CT images as well as CT scans of normal subjects.

4.3.2 Evaluation on Synthetic Dataset

The pre-traumatic photograph-based reference bone estimation is evaluated on the synthetic patient data via leave-one-out method. For each leave-one-out experiment, one normal subject is selected as the testing sample, while the remaining normal subjects are used to construct a normal facial shape database. For each testing sample, the synthetic patient data is generated as described in Section 3.1 and then fed into the framework to estimate the reference facial bone shape of the testing sample. The qualitative evaluation is completed by a different CMF surgeon, who can rank the similarity of the two surfaces using a 1–3 visual analog score (VAS, 1, the same; 2, similar but not the same; and 3, different). Figure 4.8 shows the qualitative evaluation results for eight randomly selected testing subjects, where the simulated posttraumatic, estimated reference, original normal bone surfaces, and surface distance heatmap between the estimation and original normal bone for each subject are shown in each row. The qualitative results of synthetic data also shown: the same (26/30), similar (4/30), and different (0/30). For each subject, the performance of the proposed method is

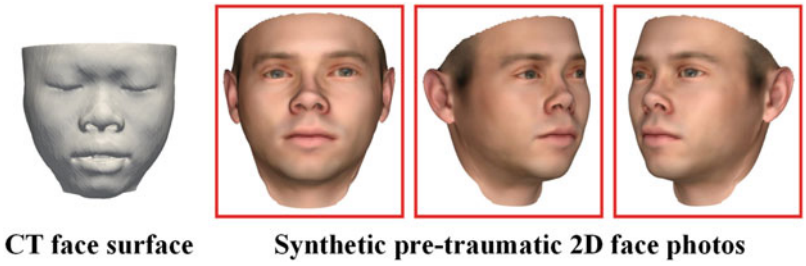


Fig. 4.7 The CT face surface of a representative normal subject and three pre-traumatic 2D face photographs simulated from the CT face surface. The figure is from Xiao et al. [19]

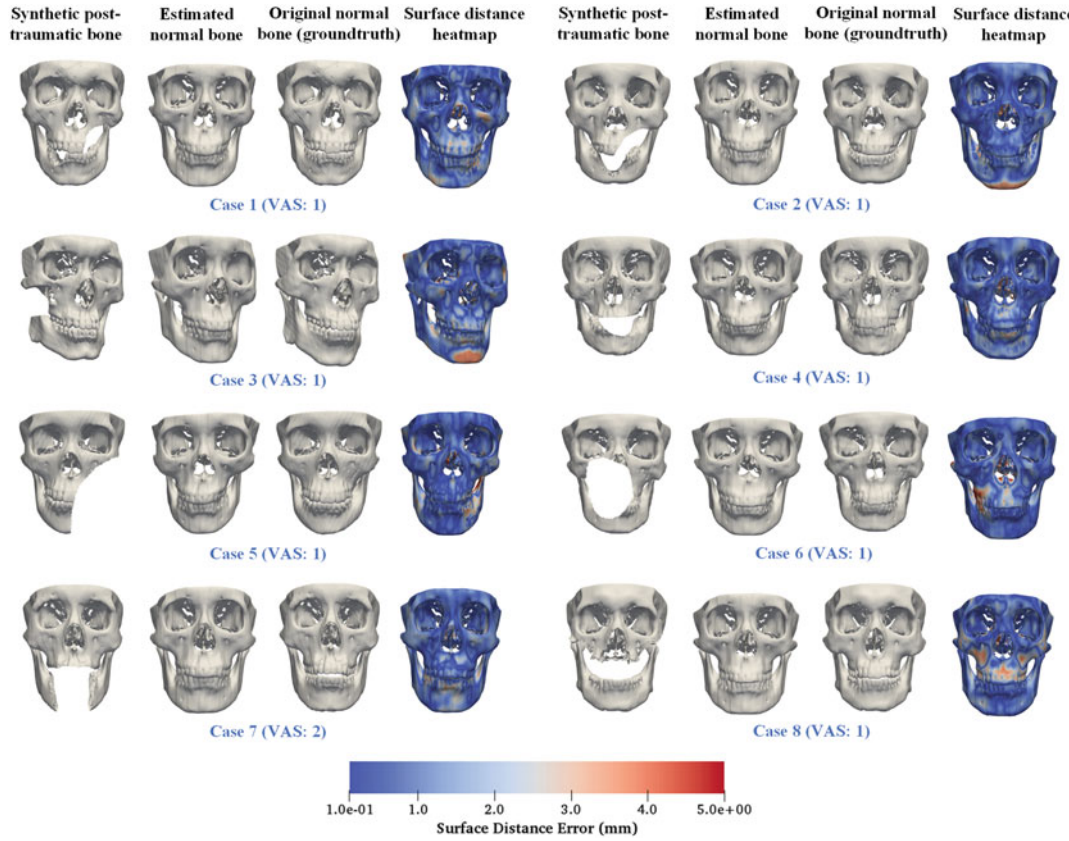


Fig. 4.8 The qualitative evaluation results of our proposed model on eight synthetic patients. The figure is from Xiao et al. [19]

Table 4.1 Facial bone estimation error (in mm) on synthetic data with the proposed approach. The table is from Xiao et al. [19]

Mean	Standard deviation	Median	Minimum	Maximum
3.68	0.43	3.66	2.91	4.90

also quantitatively evaluated by measuring an average distance error between the corresponding landmarks on the estimated and original facial bone surfaces. Results of quantitative evaluation for all cases are summarized in Table 4.1. All of these results show that the estimated reference bone shape is similar to the corresponding original normal bone shape. Thus, the defected facial bone with different types of defects can be recovered to the normal-looking shape using the proposed method.

4.3.3 Evaluation on Real Dataset

The framework is also evaluated using the real patient data. The patient's original traumatic (preoperative) bone shape, the estimated reference bone shape, and postoperative bone shape are shown in Fig. 4.9. Here the postoperative facial bone yielded by the surgery is planned using the conventional CASS method. The comparison between the estimation and postoperative bone shape is also shown in Fig. 4.9. According to an experienced CMF surgeon after visual inspection

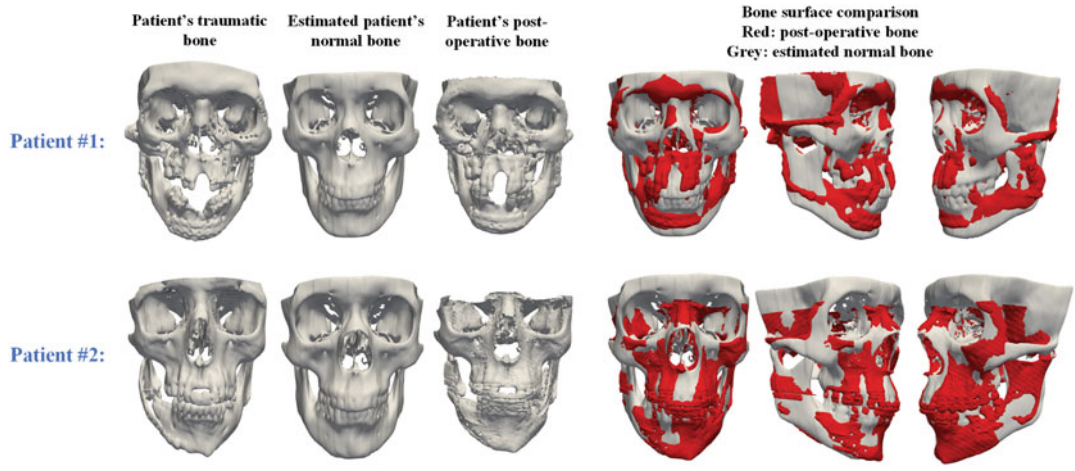


Fig. 4.9 The patient's normal facial bones estimated by applying our proposed framework on the real patient data, which are compared with patient's traumatic bones and postoperative bones after surgery. Images are taken from Xiao et al. [19]

of the results, the estimated bone shape is clinically acceptable.

For quantitative evaluation of the proposed method, an evaluation metric called recovery degree (*RD*) is defined as a product of two measurements, i.e.,

$$RD = SD \times ND, \tag{4.7}$$

where *SD* denotes the similarity degree that computes the similarity between the estimated and posttraumatic facial bones at the normal regions of the patient's posttraumatic bone, while *ND* denotes the normality degree that computes the similarity between the estimated and the statistical normal facial bone. The *SD* and *ND* are mathematically defined as

$$SD = \frac{1}{Dis\left(L_{Pat}^{normal}, L_{Est}^{normal}\right) + 1}, \tag{4.8}$$

$$\mathrm{ND} = \frac{1}{\mathrm{Dis}\left(\mathrm{L_{Recon}}, \mathrm{L_{Est}}\right) + 1}, \tag{4.9}$$

where $Dis(\bullet)$ denotes the average Euclidean distance between two sets of corresponding bone landmarks. L_{Pat}^{normal} denotes the landmarks on the normal part of the patient's posttraumatic facial bone structure, where the normal part is identified by an experienced CMF surgeon. L_{Est}^{normal} denotes the landmarks on the estimated facial bone surface corresponding to L_{Pat}^{normal}. L_{Recon} denotes the landmarks on the statistically normal facial bone that are generated by reconstructing the estimated facial bone landmarks (i.e., L_{Est}) using principal components of the normal subjects' landmarks. Specifically, the PCA technique is first performed on a matrix that contains all the facial bone landmarks of normal subjects, where each column of the matrix consists of all the vectorized landmark locations of a subject. After that, a set of principal components of landmark locations is obtained, which is subsequently used to reconstruct L_{Est} via sparse regression. The set of reconstructed landmarks using this statistical method is denoted as L_{Recon}. Up to this point, *SD* and *ND* in Eq. (4.8) and (4.9) are not properly scaled, and thus their values might be skewed and not very useful in making comparison. Therefore, the $Dis(\bullet)$ in Eq. (4.8) and (4.9) is normalized as follows:

$$\tilde{Dis} = \frac{Dis\,(\bullet) - Dis_{min}}{Dis_{max} - Dis_{min}}, \tag{4.10}$$

where $\tilde{Dis}$ means the normalized $Dis(\bullet)$, while Dis_{min} and Dis_{max} denote the minimum and the maximum of Euclidean distance errors obtained from bone landmarks on normal subjects, respectively. Specifically, for *SD*, its Dis_{min} and Dis_{max}

Table 4.2 Quantitative evaluation results for three different reference shape models generated in the proposed framework

Method	*SD*	*ND*	*RD*
Without refinement	0.367	0.999	0.366
Without SSM constraint	0.999	0.459	0.459
Proposed method	0.812	0.756	0.614

Note: *SD* denotes similarity degree, *ND* denotes normality degree, and *RD* denotes recovery degree

were set as the minimum and maximum values of the distance errors obtained in the quantitative evaluation on the synthetic patient data, respectively (i.e., $Dis^{SD}_{min} = 2.910$, $Dis^{SD}_{max} = 4.900$). Similarly, for *ND*, the minimum and maximum values were determined from another set of distance errors. The distance errors are measured on the bone landmarks of each of the normal subjects by using the same approach applied on $Dis(L_{Recon}, L_{Est})$ computation (i.e., $Dis^{ND}_{min} = 1.989$, $Dis^{ND}_{max} = 3.452$).

The recovery degrees for three different estimated reference shape models are calculated based on one real patient data. The first reference shape is generated using the proposed framework without estimation refinement. The second reference shape is generated using the proposed framework without SSM constraint in the refinement stage, and the third reference shape is generated using the proposed framework with all the components integrated. The calculated recovery degrees for these three reference shape models are listed in Table 4.2. From the comparison results in Table 4.2, it can be observed that the reference shape model generated by the complete framework achieved higher recovery degree than the other two estimated models. This reflects that both steps of estimation refinement and SSM constraint are necessary to generate an optimal estimation of normal shape.

4.4 Discussion and Conclusion

There is no definitive quantitative measure on the success of posttraumatic reconstruction due to its complexity. The current clinical standard for posttraumatic reconstruction is "surgeons do the best and patient accepts surgical outcomes." Clinicians plan surgery and evaluate postoperative outcomes subjectively for overall facial harmony, with the limited help of linear and angular measurements of size, position, orientation, and symmetry of each facial unit. Therefore, without an estimated patient-specific reference model, surgeons have to rely on their subjective evaluation as they do clinically, and this is why the reference bone estimation is important in the field of CMF skeleton reconstruction.

Compared with conventional methods for CMF skeleton reconstruction, the proposed framework can deal with more types of CMF defects. In the proposed framework, a reliable initial normal model is first estimated from pre-traumatic face photos, and then the initial estimation is refined by applying both SSM and geometric deformation (AFDSM) techniques. In this way, the advantages of both SSM and geometric deformation can be combined together, and the limitations of using solely one of them are eliminated, which eventually improves the generalization capacity of the proposed method to address various kinds of CMF defects. Similar to the initial estimation stage of the framework, sparse representation is applied in [12] to establish a correlation model between midface and jaw, which is then directly applied on the testing patient with remaining normal midface to estimate a normal jaw. Compared to [12], the estimation from the proposed correlation model is further refined with the patient's current facial bone. In this way, the estimation accuracy can be improved.

One potential issue with the proposed approach is related to the limited size of the normal subjects for training. Based on a small set of subjects, it is challenging to construct a reliable correlation model to perform initial estimation and a statistical shape model to constrain surface deformation during estimation refinement. For example, in some special cases, the facial structure of an input testing subject is obviously different from that of the reference normal subjects. Because of that, a poor initial estimation may be obtained from the correlation model and may also lead to an inaccurate bone shape with

the statistical shape model, thus resulting in a suboptimal estimation. One possible solution is to enlarge the database with more normal subjects, including both CT and MR images.

To sum up, a reference bone shape for the patient with severe CMF defects can be estimated from his/her pre-traumatic face photographs, and then the reference bone shape model can be used for guiding the CMF surgical planning. Future work on this topic can be focused on integrating the proposed method into the current surgical planning system to assist surgical planning in clinical practices. It can be expected that the clinical outcome of CMF surgery will be improved under the guidance of the updated CMF surgical planning system.

References

1. Bell RB. Computer planning and intraoperative navigation in cranio-maxillofacial surgery. Oral and Maxillofacial Surgery Clinics. 2010;22(1):135–56.
2. Xia JJ, Gateno J, Teichgraeber JF. Three-dimensional computer-aided surgical simulation for maxillofacial surgery. Atlas Oral Maxillofac Surg Clin North Am. 2005;13(1):25–39.
3. Xia JJ, Gateno J, Teichgraeber JF. New clinical protocol to evaluate craniomaxillofacial deformity and plan surgical correction. J Oral Maxillofac Surg. 2009;67(10):2093–106.
4. Xia JJ, Shevchenko L, Gateno J, Teichgraeber JF, Taylor TD, Lasky RE, English JD, Kau CH, McGrory KR. Outcome study of computer-aided surgical simulation in the treatment of patients with craniomaxillofacial deformities. J Oral Maxillofac Surg. 2011;69(7):2014–24.
5. Schmelzeisen R, Gellrich NC, Schoen R, Gutwald R, Zizelmann C, Schramm A. Navigation-aided reconstruction of medial orbital wall and floor contour in cranio-maxillofacial reconstruction. Injury. 2004;35(10):955–62.
6. Gellrich NC, Barth EL, Zizelmann C, Rüker M, Schön R, Schramm A. Computer assisted oral and maxillofacial reconstruction. J Comput Inform Technol. 2006;14(1):71–7.
7. Zachow S, Lamecker H, Elsholtz B, Stiller M. Reconstruction of mandibular dysplasia using a statistical 3D shape model. In: Proc. computer assisted radiology and surgery; 2005. p. 1238–43.
8. Semper-Hogg W, Fuessinger MA, Schwarz S, Ellis E III, Cornelius CP, Probst F, Metzger MC, Schlager S. Virtual reconstruction of midface defects using statistical shape models. J Oral Maxillofac Surg. 2017;45(4):461–6.
9. Anton FM, Steffen S, Joerg N, Carl-Peter C, Mathieu G, Philipp P, Ruediger Z, Christian MM, Stefan S. Virtual reconstruction of bilateral midfacial defects by using statistical shape modeling. J Oral Maxillofac Surg. 2019;47(7):1054–9.
10. Jolliffe IT. Principal component analysis. New York: Springer-Verlag; 1986.
11. Heimann T, Meinzer HP. Statistical shape models for 3D medical image segmentation: a review. Med Imag Anal. 2009;13(4):543–63.
12. Wang L, Ren Y, Gao Y, Tang Z, Chen KC, Li J, Shen SG, Yan J, Lee PK, Chow B, Xia JJ, Shen D. Estimating patient-specific and anatomically correct reference model for craniomaxillofacial deformity via sparse representation. Med Phys. 2015;42(10):5809–16.
13. Li Z, An L, Zhang J, Wang L, Xia JJ, Shen D. Craniomaxillofacial deformity correction via sparse representation in coherent space. In: Proc. of Mach. Learning Med. Imag; 2015. p. 69–76.
14. Xie S, Leow WK, Lim TC. Laplacian deformation with symmetry constraints for reconstruction of defective skulls. In: Proc. of International Conference on Computer Analysis of Images and Patterns; 2017. p. 24–35.
15. Bookstein FL. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans Pattern Anal Mach Intell. 1989;11(6):567–85.
16. Sorkine O, Cohen-Or D, Lipman Y, Alexa M, Rössl C, Seidel HP. Laplacian surface editing. In: Proceedings of the Eurographics/ACM SIGGRAPH Symposium on Geometry Processing; 2004. p. 179–88.
17. Xiao D, Wang L, Deng H, Thung KH, Zhu J, Yuan P, Rodrigues YL, Perez L, Crecelius CE, Gateno J, Kuang T, Shen SGF, Kim D, Alfi DM, Yap PT, Xia JJ, Shen D. Estimating reference bony shape model for personalized surgical reconstruction of posttraumatic facial defects. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention; 2019. p. 327–35.
18. Donoho DL. For most large underdetermined systems of linear equations the minimal $\ell 1$-norm solution is also the sparsest solution. Commun Pure Appl Math. 2006;59(6):797–829.
19. Xiao D, Lian C, Wang L, Deng H, Lin HY, Thung KH, Zhu J, Yuan P, Perez L, Gateno J, Shen SG, Yap PT, Xia JJ, Shen D. Estimating reference shape model for personalized surgical reconstruction of craniomaxillofacial defects. IEEE Trans Biomed Eng. 2020; https://doi.org/10.1109/TBME.2020.2990586.
20. Bulat A, Tzimiropoulos G. How far are we from solving the 2D & 3D face alignment problem?(and a dataset of 230,000 3D facial landmarks). In Proc. IEEE Int. Conf. Comp. Vis. 2017. pp. 1021–1030.
21. Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T. September. A 3D face model for pose and illumination invariant face recognition. In Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance; 2009 pp. 296–301.

22. Piotraschke M, Blanz V. Automated 3D face reconstruction from multiple images using quality measures. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2016. pp. 3418–3427.
23. Wang L, Gao Y, Shi F, Li G, Chen KC, Tang Z, Xia JJ, Shen D. Automated segmentation of dental CBCT image with prior-guided sequential random forests. Med Phys. 2016;43(1):336–46.
24. Lorensen WE, Cline HE. Marching cubes: a high resolution 3D surface construction algorithm. Computer Graphics. 1987;21(4):163–9.
25. Zhang J, Gao Y, Wang L, Tang Z, Xia JJ, Shen D. Automatic craniomaxillofacial landmark digitization via segmentation-guided partially-joint regression forest model and multiscale statistical features. IEEE Trans Biomed Eng. 2016;63(9):1820–9.
26. Myronenko A, Song X. Point set registration: coherent point drift. IEEE Trans Pattern Anal Mach Intell. 2010;32(12):2262–75.
27. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Roy Statist Soc. 2005;67(2):301–20.
28. Besl PJ, McKay ND. A method for registration of 3D shapes. IEEE Trans Pattern Anal Mach Intell. 1992;1611:586–60.
29. Shen D, Davatzikos C. An adaptive-focus deformable model using statistical and geometric information. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):906–13.
30. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vis. 1988;1(4):321–31.
31. Williams DJ, Shah M. A fast algorithm for active contours and curvature estimation. Comput Vis Graph Image Process. 1992;55(1):14–26.

Machine Learning for Facial Recognition in Orthodontics

5

Chihiro Tanikawa and Lee Chonho

5.1 Introduction

In orthodontics, the cranio-facial morphology of the patients is evaluated based on facial photographs (two-dimensional [2D] and three-dimensional [3D]), frontal and lateral cephalograms, computed tomography findings, and other resources. Facial photographs and lateral cephalograms are particularly indispensable in current clinical settings when orthodontists assess patients' cranio-facial morphology, helping physicians make an accurate diagnosis and plan appropriate treatment. Given the fact that most patients who seek orthodontic treatment anticipate improvement in their facial appearance as well as recovery of any impaired jaw or oral function, it is important for orthodontists to clearly understand the patient's facial form and its problems.

Traditionally, the major variables associated with the evaluation of the face have been the size and shape (ratios, distances, and angles). Facial forms are classified into a small number of subtypes (e.g., convex, straight, or concave type) using a few linear and angular variables that are measured fractionally. This classification method is simple and convenient but has difficulty expressing complex and holistic information concerning faces [1].

The face is an important site for the identification of others, and it also conveys significant social information. The face generally is considered to be an interface that has the capability to transmit rich information including races, sex, age, and facial expressions [2]. Human experts, such as orthodontists and dysmorphologists, can instantly evaluate the facial appearance on first sight and often detect even specific syndromes or health problems [3]. This is based on the fact that the human brain has a series of perception systems specific to faces (facial recognition in the brain is briefly explained in the next section [4]).

Recent technological advances have facilitated access to AI technology, which has resulted in

C. Tanikawa (✉)
Graduate School of Dentistry, Osaka University, Osaka, Japan

Center for Advanced Medical Engineering and Informatics, Osaka University, Osaka, Japan
e-mail: ctanika@dent.osaka-u.ac.jp

L. Chonho
Cybermedia Center, Osaka University, Osaka, Japan

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_5

an increase in the application of AI to facial recognition. For example, Tanikawa et al. developed an automated phenotyping system of holistic facial profiles including the lips and noses for orthodontic evaluation purposes [5–7] using a pattern matching algorithm. Pattern matching is a computer process that simulates human cognition (for details of this pattern matching system, please see the section below). Their studies showed that feature extraction processing incorporating the knowledge and experience of human experts was more effective in creating facial profile patterns than simple mechanical sectioning.

A recent breakthrough in image recognition technology using the deep convolutional neural network (DCNN) model [8, 9] brought about dramatic improvement in diagnostic imaging. The DCNN is a system developed based on the basic model of the brain function in cognitive science established in 1951 [10] with the goal of creating a "more generalized system" similar to the human brain. DCNN methods have been applied to facial recognition. Recently, a facial recognition software program (Face2Gene [11]) was developed to detect genetic problems as a starting point in cases where a doctor does not know what to make of a patient's symptoms. Facial phenotyping using this system can provide a genetic diagnosis, even in patients with small, supernumerary marker chromosomes [12]. More recently, systems that automatically provide clinical descriptions of oral or facial images for orthodontic diagnostic purposes have been reported [13, 14]. Such systems provide additional objective measurements and data to confirm the descriptive features for facial phenotyping. These AI systems can provide an objective facial morphological assessment by identifying clinically useful facial traits during the orthodontic diagnosis process (e.g., concave profile, upper lip retrusion, presence of scars). This automation considerably reduces the assessment workload for dentists and also prevents variations in the diagnosis. Therefore, in the first half of this chapter, we will briefly discuss this algorithm.

In the last half of this chapter, we will introduce an automated system for analyzing craniofacial morphology using cephalograms. Analyzing cephalograms and thereby clarifying a patient's cranio-facial morphological characteristics are one of the most important steps in the orthodontic diagnosis. There are three steps to this analysis: identification of the landmarks, measurement of the variables (angles and distances), and the comparison of these variables with normal ranges. The last two steps were first automated in the 1980s [15]. However, the first step (landmark identification) has been challenging from the perspective of computer engineering.

Because a fully automated cephalometric analysis will significantly reduce experts' workload and provide a more accurate analysis, various AI systems and methods for automating landmark identification have been proposed, such as image processing approaches [15–17], model-based approaches [18–20], combined approaches [21–23], and machine learning-based approaches [24–26]. However, these approaches require a large number of cephalograms to improve the accuracy, and identification performance has not proven to be very satisfactory in practical use.

The DCNN model has brought about dramatic improvement in diagnostic imaging. This methodology can be used to automatically diagnose the presence of tuberculosis in chest radiographs [27], to detect macular degeneration from fundus images [28], to locate malignant melanoma in skin images [29], and automatic recognition of cephalograms [30, 31]. Especially, the work [31] employed two deep neural networks and succeeded in identifying the landmarks with practical accuracy. That system was trained by multi-scale patches, i.e., cropped images that contain landmarks, whose rectangle size varies based on landmark-dependent criteria [32]. In the last half chapter, we will introduce this system in detail.

In the final section, we will discuss the challenges associated with deep learning in medical imaging. The number of training datasets can be problematic when developing a DCNN [33]. The more complex the tasks for AI, the greater the number of datasets needed. However, it is difficult to obtain a large number of datasets for medical images, especially for rare cases. Incorporating an expert thinking process when creating an AI sys-

tem can help mitigate these limitations associated with medical images. Understanding the thinking of experts concerning facial recognition may help create AI systems that require only a minimum number of datasets.

5.2 Background Knowledge in Cranio-facial Recognition

5.2.1 Facial Perception and Recognition by the Brain

Facial perception is considered to be subserved by areas of the facial processing network in the ventral stream of the brain, i.e., the occipital face area (OFA [34]), fusiform face area (FFA [35]), and ventral anterior temporal lobe (vATL [36]). The OFA is primarily sensitive to low-level perceptual attributes of faces, such as the viewpoint [37] and location. The FFA is primarily involved in the holistic processing of faces and responds to the shape of facial features as well as the spacing between them [38]. The right ATL is typically activated during the discrimination of familiar and unfamiliar faces [39] and face naming [40]. In addition, multivoxel activity patterns in the vATLs discriminate between individual facial identities [41]. While the OFA, FFA, and AT face area of the brain (ventral stream) are primarily involved in processing face identity, the posterior superior temporal sulcus (pSTS) in the temporal lobe of the brain (dorsal stream) has been implicated in processing facial expression and eye gaze, information that is derived from the moveable or dynamic aspects of a face [4].

Based on these findings, human facial perception and recognition process can be modeled [4, 42]. The core system consists of (1) detecting the presence of a face (OFA); (2) extracting features as well as the spacing between them (FFA); and (3) processing dynamic features, such as the facial expression, gaze, and lip movement (STS). On the other hand, the extended system consists of regions that are also parts of neural systems for other cognitive functions, i.e., the AT for recognition of the face in terms of identity (identity, name, biography) and the amygdala, insula, or limbic system for the recognition of the face regarding emotions [4].

Computer vision systems may employ a similar process [43]. Recently, a study comparing human brain activity with that of a DCNN revealed that human performance for facial recognition significantly matched that of the DCNN [44]. The authors further examined whether or not the face-space geometry of face exemplars, as determined by pair-wise distances between their activation patterns, was similar between the human cortex and individual DCNN layers. They concluded the importance of face-space geometry in the pictorial aspects of human face perception, which are similar to the mid-hierarchical layers of the DCNN. As such, DCNN well represents human perceptions, and making AI systems can help clarify the human recognition process.

5.2.2 Pattern Recognition

In psychology and cognitive neuroscience, pattern recognition is considered to be "the process whereby environmental stimuli are recognized as exemplars of concepts and principles already in memory." Pattern recognition represents a fundamental aspect of human cognition. In the field of AI, pattern recognition basically indicates the use of automation to help identify the regularities of patterns in data through the use of computer algorithms [45]. Pattern recognition mainly involves three processes: data building, pattern analysis, and pattern classification. Data building is the conversion of original information into a vector that can be dealt with by a computer. Pattern analysis is the processing of data (vector) by feature selection, data-dimension compression, and regression, which helps detect regularities in order to classify data. Pattern classification aims to utilize the information acquired from pattern making to discipline the computer in order to accomplish classification, assigning an object to a class.

The output of the pattern recognition system is an integer label, such as classifying a product as "1" or "0" in a quality control test [46].

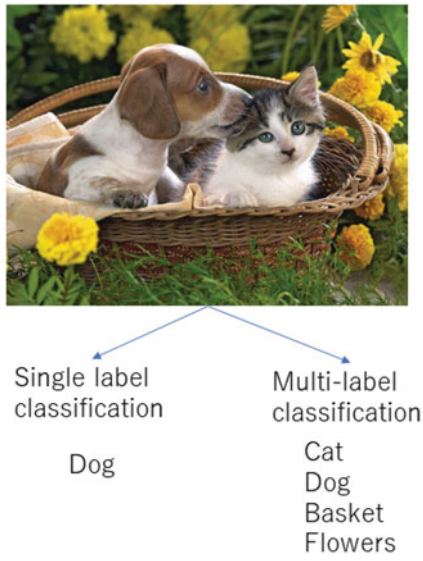


Fig. 5.1 An example of multi-label classification

The primary goal of pattern recognition is supervised or unsupervised classification. Recently, neural network techniques (or deep learning) and methods derived from statistical approaches (e.g., support vector machine) have received increasing attention.

5.2.3 Multi-label Image Classification

In machine learning, multi-label classification is a variant of the classification approach where multiple labels may be assigned to each image (Fig. 5.1). Generally, each medical image will have several points of focus, so multi-label classification is useful.

5.2.4 Convolutional and Recurrent Neural Networks

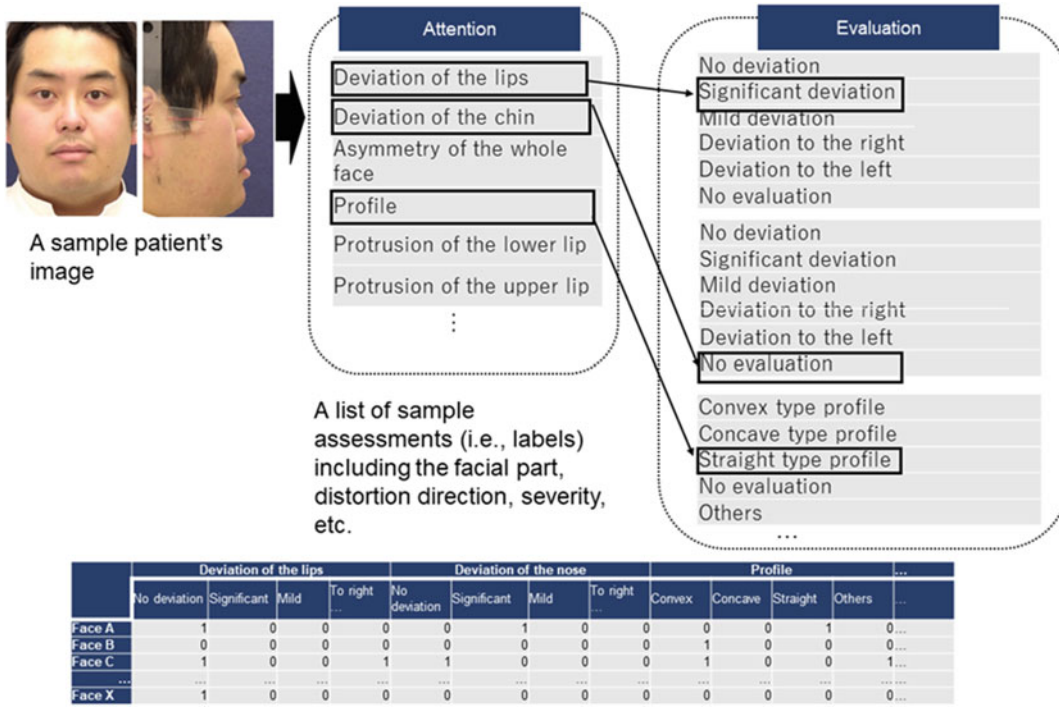
In general, there are two traditional models for the deep learning of images: the convolutional neural network (CNN) and the recurrent neural network (RNN). The CNN is a traditional neural network model that is generally composed of convolutional layers (where filters extract the target features), pooling layers (where the spatial sizes of features and the amount of model parameters are reduced), and fully connected layers (a linear combination of the features of the previous layer that makes the next layer). While the CNN is a feed-forward neural network that is generally used for image recognition and classification, the RNN works on the principle of saving the output of a layer and feeding this back in as input to predict the output of the layer, which is particularly useful when a sequence of data is being processed to make a classification decision, such as with time-series data.

An orthodontic facial diagnosis is also a time-series examination, as orthodontists diagnose a patient comprehensively by looking at the entire face while assessing multiple parts of the face from different angles, rather than by simply targeting one part of the face. For instance, an orthodontist must first look at the frontal plane of a patient's face and search for any asymmetry, including the inclination of the eyelids and/or distortion of the nose. The orthodontist must also check for the presence of maxillary protrusion and/or prognathism from the side of the face, check the tooth alignment while smiling, and finally give the patient a facial diagnosis. This complex assessment process consequently results in variation in the diagnosis of different orthodontists. To mimic an orthodontist's comprehensive process using an AI system, Murata et al. [14] employed an RNN with the attention mechanism, as detailed below.

5.3 Systems that Automatically Provide Clinical Descriptions of Facial Images for Orthodontic Diagnostic Purposes

5.3.1 Sample Data

Lateral and frontal facial images of patients who visited the orthodontic department (1000 patients; male = 397; female = 603; age range 5–68 years old) were used as the training and evaluation datasets. An experienced orthodontist examined



	Deviation of the lips				Deviation of the nose				Profile				...
	No deviation	Significant	Mild	To right ...	No deviation	Significant	Mild	To right ...	Convex	Concave	Straight	Others	...
Face A	1	0	0	0	0	1	0	0	0	0	1	0	...
Face B	0	0	0	0	0	0	0	0	1	0	0	0	...
Face C	1	0	0	1	1	0	0	0	1	0	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Face X	1	0	0	0	0	0	0	0	0	0	0	0	...

Fig. 5.2 Sample patient's image and list of sample assessments (i.e., labels), including the region of interest, evaluation findings, etc.

all of the facial images for each patient and identified as many facial traits that would be clinically useful during the orthodontic diagnosis process as possible (e.g., deviation of the lips, deviation of the mouth, asymmetry of the face, concave profile, upper lip retrusion, presence of scars). A sample patient's image, a list of sample assessments (i.e., labels), and the multi-label data used are shown in Fig. 5.2. A total of 161 clinical anatomical traits concerning 35 regions of interest (ROIs) (Fig. 5.2) were extracted. The average number of traits was 6.5 and ranged from 1 to 18.

5.3.2 Model Multi-label Image Classification

In general, medical images may contain multiple RIOs to be evaluated. An orthodontist diagnoses a patient based on the assessment results of various facial regions from several different facial images. Therefore, for automated diagnostic imaging, a typical single-label (binary or multiclass) image classification model was used to resolve the issue of multi-label image classification.

5.3.3 Results

Lateral and frontal images and the orthodontist's multiple answers were used as the dataset for each patient. One hundred datasets were reserved for the system test, while the other 900 were used for system development. To evaluate the performance reliability, the system examined the reserved datasets. As a result, the system successfully identified all specified anatomical traits in all images. The success rate was 95%, and the systems detect at least one correct clinical anatomical trait in 91% of the test datasets.The

Table 5.1 The definition of the accuracy, sensitivity, precision, and specificity used to evaluate the system

	Traits labeled 1 by the orthodontist	Traits labeled 0 by the orthodontist
Traits classified as 1 by the AI	A	B
Traits classified as 0 by the AI	C	D

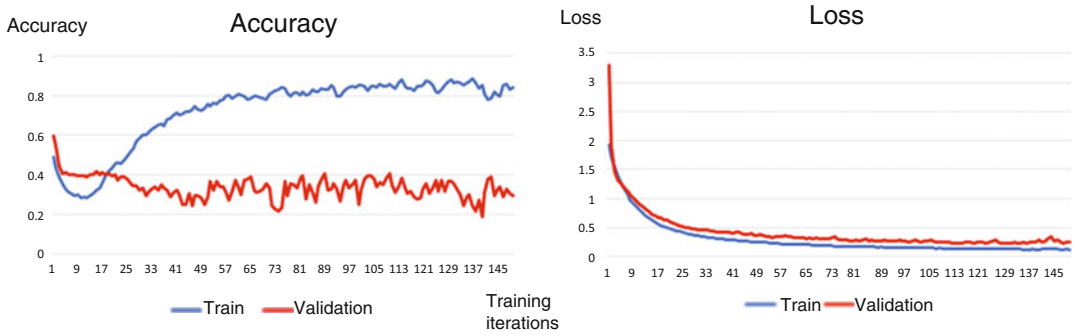


Fig. 5.3 Learning curve

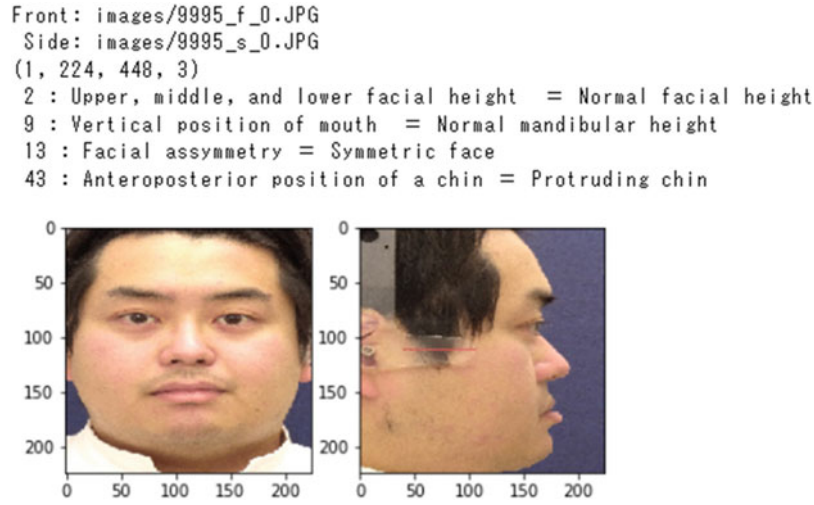


Fig. 5.4 Examples of the system response. Four correct descriptions were detected, but certain traits, such as a small nasolabial angle and the location of the right corner of the mouth in a high position, were missed

performance results are shown in Table 5.1. The learning curves and example system response are shown in Figs. 5.3 and 5.4, respectively.

Each pair of frontal and lateral facial images was evaluated for 161 labels. If the orthodontist considered a pair of images to have a certain trait, the trait was labeled as 1; otherwise, it was labeled as 0. If the AI system detected a certain trait in a pair of frontal and lateral facial images, the trait was classified as 1; otherwise, it was classifiedas 0. A is the number of traits labeled 1 by the orthodontist and classified as 1 by the AI; B is the number of traits labeled 0 by the orthodontist and classified as 1 by the AI; C is the number of traits labeled 1 by the orthodontist and classified as 0 by the AI; D is the number of traits labeled 0 by the orthodontist and classified as 0 by the AI. Accuracy = (A + D)/(A + B + C + D) [%]; sensitivity or recall = A/(A + C) [%]; precision = A/(A + B) [%]; specificity = D/(B + D) [%].

Table 5.2 Total number of traits labeled by the orthodontist and classified by the AI system

	Traits labeled 1 by the orthodontist	Traits labeled 0 by the orthodontist
Traits classified as 1 by AI	893	1595
Traits classified as 0 by AI	1411	53,256

The accuracy of the AI system was 95%, sensitivity was 39%, precision was 36%, and specificity was 97% (Table 5.2).

5.4 Deep Learning-Based Cephalometric Landmark Identification Using Landmark-Dependent Multi-scale Patches

5.4.1 Dataset and Method

This section introduces a cephalometric landmarking model using deep neural networks trained by multi-scale image patches. The following two subsections describe two different phases: the training phase and landmarking phase. Two deep neural networks were used for landmark patch classification and landmark point estimation. In addition, how landmarks in a given cephalogram are estimated using the trained neural networks is also described.

5.4.1.1 Training Phase

Dataset

A dataset was created based on cephalograms of 936 patients (476 men and 460 women), provided by the orthodontic department of Osaka University Dental Hospital. The mean age was 10.98 ± 3.57 (range: 4–32) years old. The original cephalogram size was 2100 × 2500 pixels with a pixel spacing of 0.1 mm. As shown in Fig. 5.5, the landmark points (i.e., true values) of 22 hard tissues and 11 soft tissues were plotted by one orthodontist (4 years' experience). Another orthodontist (19 years' experience) then checked these plotted landmarks. The landmark data were used as the ground truth values.

The dataset contains multi-scale image patches, including landmarks, with their name and location. The landmark name and location were used as labels during training of the neural networks for patch classification and point estimation, respectively. Patches with various rectangle shapes were extracted, and the point location was stored in the patch as a set of x and y coordinates. These coordinates were not derived from the original cephalogram. For example, the multi-scale patches were extracted so that the patches included the several surrounding anatomical features. For the nasion, the frontal sinus, nasal bone, and eyelid were cropped and employed as patches for recognition. This multi-scale approach simulates the identification process used by orthodontists. From among 935 cephalograms, 9000 patches per landmark (i.e., 10 patches per cephalogram) were extracted in total. In addition, data augmentation operations were performed for the patches during training by rotation and changing the gamma values. The structure of the deep neural networks was as follows: CNN (named CNN-PC) consisting of a set of three convolution and pooling layers and two fully connected layers with a softmax layer.

5.4.1.2 Landmarking Phase

In this phase, landmark points for a given cephalogram are estimated using the trained neural networks. This phase consists of two steps: patch classification and point estimation. First, grid points are set on a cephalogram. For each grid point, multiple patches are cropped using multi-scale windows and classified by the CNN-PC. The classified and resized patches are then stored as candidate patches of corresponding landmarks. Second, given a set of candidate patches, x and y coordinates are estimated by checking the distribution of the scatter plot of the landmark candidates. Based on the scatter plot, one mean/median point is computed as a landmark point, followed by the omission of outliers for points with a Euclidean distance

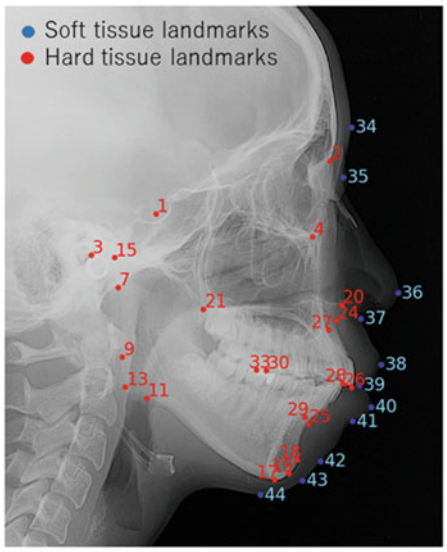
from the mean/median greater than two times the standard deviation (i.e., $\pm 2\sigma$). One interesting point here is that this landmark estimation process also simulates the orthodontists' identification process. Orthodontists generally identify several candidates from among several anatomical features and combine this information with their innate knowledge, such as the nasion being located below the frontal sinus but above the eyelid and nasal bone.

5.4.2 Evaluations

All experiments were performed in Ubuntu 16.04 LTS with a quad core 3.6-GHz CPU, 128-GB RAM, and Nvidia P100 GPU. Training the proposed CNN models took 10 s per epoch for 128 batches. The CNN-PC achieved around 93% validation accuracy and 0.4 validation loss in 1000 epochs. In testing, it takes about 0.7 s to identify one of the major landmarks shown in Fig. 5.5. At Osaka University Dental Hospital, this approach dramatically reduced the manual landmarking time, as orthodontists spent about 5 to 7 min identifying 30 major landmarks.

5.5 Results

Two deep neural networks are trained by multi-scale patches, i.e., cropped images that contain landmarks, whose rectangle size varies based on landmark-dependent criteria examined by orthodontists, and 22 hard-tissue and 11 soft-tissue landmarks are identified. For the evaluation, (i) the landmark estimation accuracy based on the Euclidean distance error between true and estimated values and (ii) the success rate for an estimated landmark being located within the corresponding norm using a confidence ellipse are computed. The proposed model successfully identified hard-tissue landmarks with an error range of 1.32–3.5 mm and a mean success rate of 96.4% and soft-tissue landmarks with an error range of 1.16–4.37 mm and a mean success rate of 75.2%. For testing, this approach would dramatically reduce the landmarking time for 33 landmarks from about 5 to 7 min for an orthodontist to 21 s for the network.



1: S (Sella)
2: N (Nasion)
3: Po (Porion)
4: Or (Orbitale)
7: Ar (Articulare)
9: Sup.go (Sup. gonion)
11: Inf.go (Inf. gonion)
13: Go (Gonion)
15: Cd (Condylion)
17: Me (Menton)
18: Pog (Pogonion)
19: Gn (Gnathion)
20: Ans (Ant. nasal spine)
21: Pns (Post. nasal spine)
24: Point A (A point)
25: Point B (B point)
26: U1 (Upper incisor)
27: U1_a (Upper incisor root apex)
28: L1 (Lower incisor)
29: L1_a(Lower incisor root apex)
30: U6 (Upper molar)
33: L6 (Lower molar)
34: Gla (Glabella)
35: N' (Soft tissue nasion)
36: Prn (Pronasale)
37: Sn (Subnasale)
38: Ls (Labialis superior)
39: Sto (Stomion)
40: Li (Labial inferior)
41: Sm (Submentale)
42: Pog' (Soft tissue pogonion)
43: Gn' (Soft tissue gnathion)
44: Me' (Soft tissue menton)

Fig. 5.5 Landmark point

5.6 Discussions for Recognition of Medical Facial Images

We consider there to be three main challenges that must be addressed concerning the creation of deep learning systems that recognize medical facial images.

First, deep learning requires a massive training dataset in order to obtain accurate classifications. However, the availability of medical images is quite limited, and the annotation of new datasets is extremely time-consuming for medical experts. Furthermore, for rare cases, a sufficient number of datasets to accurately reflect the medical status may be difficult to obtain. In this chapter, the first systems had 161 clinical anatomical traits concerning 35 ROIs. Of these, several anatomical features were present in only 10 of 1000 patients, hampering learning. In the future, we will address this problem by increasing the numbers of datasets and mathematical approaches.

Second, the quality of the dataset and annotation can be an issue. To control the quality of the data, the introduced systems used only data from a single hospital, and annotation was conducted by a few orthodontists. If the datasets are collected from multiple hospitals, the quality of the data should be standardized. Annotation is a tedious task for medical experts and sometimes generates systematic and random errors. Systematic errors can be derived from the educational system where experts learn. This issue is quite difficult to eliminate.

The third issue is a "black-box" problem for the deep learning. Conventional machine learning is suitable for understanding and interpreting the extracted features while also carefully weighing all feature variables. However, the deep learning model has difficulty interpreting weight and extracted features. For example, Face2Gene detects genetic problems, but it provides no reason as to why the system recognizes it. As medicine is a field in which reasons are required for medical intervention, another system to explain the systems' answer is needed. These newly introduced facial recognition systems may function as explanatory systems to complement holistic classification systems (e.g., Face2Gene) by describing the facial characteristics in the future.

5.7 Conclusions

This chapter introduced two AI systems that can be used to analyze cranio-facial morphology in orthodontics. Experts' knowledge and experience were incorporated into these two systems. However, while the second automated cephalometric system showed a high performance and reached a clinically applicable level, issues associated with the number of datasets, which are specific issues associated with AI of medical images, need to be resolved in the first facial description system before it can be introduced clinically.

References

1. Takada K. Elements of orthodontics. Medigit: Osaka; 2010.
2. Perrett DI, Lee KJ, Penton-Voak I, Rowland D, Yoshikawa S, Burt DM, Henzi SP, Castles DL, Akamatsu S. Effects of sexual dimorphism on facial attractiveness. Nature. 1998;394(6696):884–7.
3. Chen W, Qian W, Wu G, Chen W, Xian B, Chen X, Cao Y, Green CD, Zhao F, Tang K, Han JD. Three-dimensional human facial morphologies as robust aging markers. Cell Res. 2015;25(5):574–87. https://doi.org/10.1038/cr.2015.36.
4. Haxby JV, Hoffman EA, Gobbini MI. The distributed human neural system for face perception. Trends Cogn Sci. 2000;4(6):223–33.
5. Tanikawa C, Kakiuchi Y, Yagi M, Miyata K, Takada K. Knowledge-dependent pattern classification of human nasal profiles. Angle Orthod. 2007;77(5):821–30.
6. Tanikawa C, Nakamura K, Yagi M, Takada K. Lip vermilion profile patterns and corresponding dentoskeletal forms in female adults. Angle Orthod. 2009;79(5):849–58.
7. Tanikawa C, Takada K. Objective classification of nose-lip-chin profiles and their relation to dentoskeletal traits. Orthod Craniofac Res. 2014;17(4):226–38.
8. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst. 2012;1
9. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang L, Wang G, Cai J, Chen T. Recent advances in convolutional neural networks. arXiv preprint arXiv. 2017;1512:07108.

10. Minsky M. A neural-analogue calculator based upon a probability model of reinforcement. Psychological Laboratories, Harvard University: Cambridge, MA; 1952.
11. Face2Gene. https://www.face2gene.com/
12. Liehr T, Acquarola N, Pyle K, St-Pierre S, Rinholm M, Bar O, Wilhelm K, Schreyer I. Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. Clin Genet. 2018;93(2):378–81. https://doi.org/10.1111/cge.13087.
13. Murata S, Ishigaki K, Lee C, Tanikawa C, Date S, Yoshikawa T. Towards a smart dental healthcare: an automated assessment of orthodontic treatment need. 2017;(c):35–39.
14. Murata S, Lee C, Tanikawa C, Date S. Towards a fully automated diagnostic system for orthodontic treatment in dentistry. In: 2017 IEEE 13th International Conference on E-Science. 2017:1–8. https://doi.org/10.1109/eScience.2017.12
15. Richardson A. A comparison of traditional and computerized methods of cephalometric analysis. Eur J Orthod. 1981;3(1):15–20.
16. Rudolph DJ, Sinclair PM, Coggins JM. Automatic computerized radiographic identification of cephalometric landmarks. Am J Orthod Dentofac Orthop. 1998;113(2):173–9.
17. Grau V, Alcañiz M, Juan MC, Monserrat C, Knoll C. Automatic localization of cephalometric landmarks. J Biomed Inform. 2001;34(3):146–56.
18. Forsyth DB, Davis DN. Assessment of an automated cephalometric analysis system. Eur J Orthod. 1996;18(5):471–8.
19. Saad AA, El-Bialy AM, Ahmed A. Automatic cephalometric analysis using active appearance model and simulated annealing. ICGST Int J Graph Vis Image Process Spec Issue Image Retr Represent. 2006;6:253–7.
20. Yue W, Yin D, Wang G, Xu T. Automated 2-D cephalometric analysis on X-ray images by a model-based approach. IEEE Trans Biomedical Eng. 2006;53:1615–23.
21. Kafieh R, Mehri A, Sadri S. Automatic landmark detection in cephalometry using a modified active shape model with sub image matching. In Proc. of IEEE International Conference on Machine Vision. 2008;73–78.
22. Keustermans J, Mollemans W, Vandermeulen D, Suetens P. Automated cephalometric landmark identification using shape and local appearance models. In: Proc. of the 20th IEEE Conference on Computer Vision and Pattern Recognition; 2010.
23. Vucinić P, Trpovski Z, Sćepan I. Automatic landmarking of cephalograms using active appearance models. Eur J Orthod. 2010;32(3):233–41. https://doi.org/10.1093/ejo/cjp099.
24. Chakrabartty S, Yagi M, Shibata T, Cauwenberghs G. Robust cephalometric identification using support vector machines. In Proc. of International Conference on Multimedia and Expo 2003.
25. Giordano D, Leonardi R, Maiorana F, Cristaldi G, Distefano ML. Automatic landmarking of cephalograms by cellular neural networks. Artif Intell Med. 2005:333–42.
26. Leonardi R, Giordano D, Maiorana F. An evaluation of cellular neural networks for the automatic identification of cephalometric landmarks on digital images. J Biomed Biotechnol. 2009;2009:717102. https://doi.org/10.1155/2009/717102.
27. Ting DS, Yi PH, Hui F. Clinical applicability of deep learning system in detecting tuberculosis with chest radiography. Radiology. 2018;286(2):729.
28. Burlina PM, Joshi N, Pekala M, Pacheco KD, Freund DE, Bressler NM. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA ophthalmology. 2017;135(11):1170–6.
29. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115.
30. Arık SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham). 2017;4(1):014501. https://doi.org/10.1117/1.JMI.4.1.014501.
31. Lee CH, Tanikawa C, Lim JY, Yamashiro T. Deep learning based cephalometric landmark identification using landmark-dependent multi-scale patches. Arxiv 2019. http://arxiv.org/abs/arXiv:1906.02961
32. Tanikawa C, Yagi M, Takada K. Automated cephalometry: system performance reliability using landmark-dependent criteria. Angle Orthod. 2009;79(6):1037–46. https://doi.org/10.2319/092908-508R.1.
33. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB. Convolutional neural networks for medical ImageAnalysis: full training or fine tuning? Arxiv. 2017. https://arxiv.org/pdf/1706.00712.pdf
34. Fairhall SL, Ishai A. Effective connectivity within the distributed cortical network for face perception. Cereb Cortex. 2007;17(10):2400–6. https://doi.org/10.1093/cercor/bhl148.
35. Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci. 1997;17:4302–11.
36. Rajimehr R, Young JC, Tootell RBH. An anterior temporal face patch in human cortex, predicted by macaque maps. Proc Natl Acad Sci U S A. 2009;106(6):1995–2000. https://doi.org/10.1073/pnas.0807304106.

37. Ewbank MP, Andrews TJ. Differential sensitivity for viewpoint between familiar and unfamiliar faces in human visual cortex. NeuroImage. 2008;40(4):1857–70.
38. Liu J, Harris A, Kanwisher N. Perception of face parts and face configurations: an fMRI study. J Cogn Neurosci. 2010;22(1):203–11. https://doi.org/10.1162/jocn.2009.21203.
39. Nakamura K, Kawashima R, Sato N, Nakamura A, Sugiura M, Kato T, Zilles K. Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study Brain. 2000;123(9):1903–12.
40. Grabowski TJ, Damasio H, Tranel D, Ponto LL, Hichwa RD, Damasio AR. A role for left temporal pole in the retrieval of words for unique entities. Hum Brain Mapp. 2001;13(4):199–212.
41. Kriegeskorte N, Formisano E, Sorger B, Goebel R. Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A. 2007;104:20600–5. https://doi.org/10.1073/pnas.0705654104.
42. Atkinson AP, Adolphs R. The neuropsychology of face perception: beyond simple dissociations and functional selectivity. Philos Trans R Soc Lond Ser B Biol Sci. 2011 Jun 12;366(1571):1726–38. https://doi.org/10.1098/rstb.2010.0349.
43. Tsao DY, Livingstone MS. Mechanisms of face perception. Annu Rev Neurosci. 2008;31:411–37. https://doi.org/10.1146/annurev.neuro.30.051606.094238.
44. Grossman S, Gaziv G, Yeagle EM, et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. Nat Commun. 2019;10:4934. https://doi.org/10.1038/s41467-019-12623-6
45. Srihari SN. Covindaraju, pattern recognition. London: Chapman & Hall; 1993. p. 1034–41.
46. Liu J, Sun J, Wang S. Pattern recognition: an overview. IJCSNS International Journal of Computer Science and Network Security. 2006;6(6):57–61.

Part II

Machine Learning for Oral Diagnosis and Treatment

6 Machine/Deep Learning for Performing Orthodontic Diagnoses and Treatment Planning

Chihiro Tanikawa, Tomoyuki Kajiwara, Yuujin Shimizu, Takashi Yamashiro, Chenhui Chu, and Hajime Nagahara

6.1 Introduction

An orthodontic diagnosis and treatment plan involve predicting the entire course of action that a dentist should take to obtain the optimal treatment results at the lowest possible risk. Making such an assessment requires years of knowledge and experience. As such, there are cases in which inexperienced dentists make judgment errors or otherwise misunderstand a case's parameters. Usually the orthodontist will not be satisfied with the analysis alone because the purpose of a diagnosis is not only to analyze the nature of the patients but also to integrate the pieces of information that have been obtained through each examination in order to understand the patient holistically [1]. This integrated knowledge will provide an accurate assessment of the true nature of the patient's dental, facial, and sociopsychological condition [1].

In modern medicine, this integrated process of making a diagnosis and planning the treatment is generally conducted through a problem-oriented approach [2]. Problem-oriented approaches involve three steps: (1) the collection of a database of information about the patient, (2) the development of a comprehensive list of the patient's problem, and (3) the specification of the treatment strategy that would provide the maximum benefit for this particular patient (Fig. 6.1). Furthermore, treatment operates on the logic of "doing the opposite of the problem." Mathematically, this can be considered to be an "optimization problem," which is an issue of finding the best solution among all feasible solutions. In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically weighting or choosing input variables from within an allowed set and computing the value of the function. In other words, putting the diagnosis process of an expert in mathematical terms, if patient information regarding the problem is thought of as a set of feature values, the above step (1) is analogous to collecting the

C. Tanikawa · Y. Shimizu · T. Yamashiro
Graduate School of Dentistry, Osaka University, Osaka, Japan

C. Tanikawa (✉)
Center for Advanced Medical Engineering and Informatics, Osaka University, Osaka, Japan
e-mail: ctanika@dent.osaka-u.ac.jp

T. Kajiwara · C. Chu · H. Nagahara
Institute for Datability Science, Osaka University, Osaka, Japan

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_6

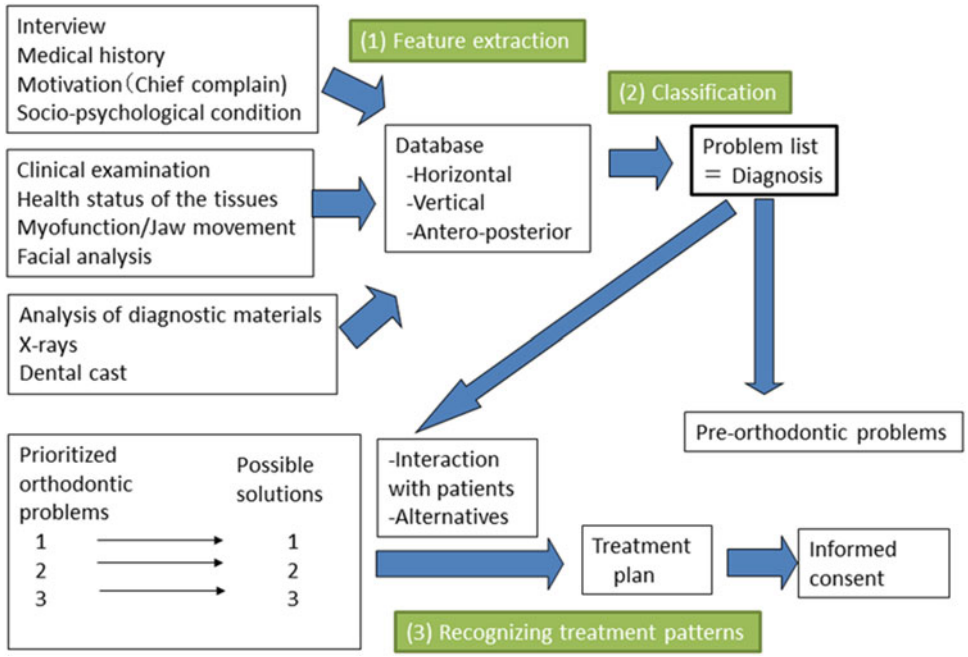


Fig. 6.1 The problem-oriented approach to the diagnosis and corresponding AI processes

feature elements to represent a patient's status as a feature vector (feature extraction). Step (2) can be considered equivalent to representing medical conditions based on the individual weight of each feature value and detecting the degree of similarity between each medical condition (data-dimension compress or classification). Step (3) can then be considered equivalent to recognizing the treatment patterns based on the stored cases or experiences that maximized the patient's benefit at the lowest possible risk.

A certain regularity is apparent in the logical structure involved in the medical diagnosis and treatment planning (as described above). Attempts have therefore been made to automate the orthodontic diagnosis and treatment planning, such as through the use of expert systems (e.g., an orthodontic diagnosis support system using fuzzy logic). However, support systems for making an orthodontic diagnosis and planning treatment in clinics have yet to be established.

In this chapter, we briefly review the previously developed automated systems for making the diagnosis and planning treatment in orthodontics. We also describe a newly developed AI system that uses natural language processing (NLP) to conduct various clinical text evaluations and develop accompanying treatment protocols.

6.2 Various AI Systems for Determining Treatment Plans in Orthodontics

In 1959, *rule-based expert systems were developed in the fields of knowledge engineering*, where the knowledge of human experts was modeled and incorporated inside the system. One of the earliest medical expert systems was the MYCIN [3], which advised physicians about the selection of antimicrobial therapy, with the dosage adjusted for the patient's body weight. These expert systems modeled human knowledge by "production rules," which is a framework for expressing the knowledge that "if a certain condition is met, a certain action is performed (IF-THEN method) [4]."

In orthodontics, several expert systems have been developed. Sims-Williams et al. [5] first developed and introduced an expert system for

orthodontics that used fuzzy logic to implement IF-THEN rules. Fuzzy logic was developed in 1965 [6] to handle the concept of partial truth, where the truth value may range between completely true and completely false. In fuzzy logic, IF-THEN rules are employed with a membership function, which is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. In other words, fuzzy logic is a multivalued logic that allows intermediate values of certainty to be defined between conventional deterministic two-valued logic such as yes/no, high/low, or true/false.

Sims-Williams et al. [5] generated a fuzzy model expert system that models the advice that orthodontic consultants give to general dental practitioners for the alignment of crowded lower teeth, especially extraction of the lower premolars. In their study, they employed three membership functions, including the determination of the degree of crowding, because fuzzy logic was considered a highly suitable and applicable basis for developing knowledge-based systems in medicine for tasks such as the interpretation of sets of orthodontic findings. Several applications to orthodontic tasks have been reported, such as the development of an expert system for treatment planning of Class II Division 1 cases [7, 8]; a fuzzy expert system to classify Class I, II, and III by implementing the orthodontists knowledge [9]; and a system for head gear-type selection [10]. As an alternative to the fuzzy logic model, Poon et al. [11] employed the Ripple-Down Rules method, which was an incremental approach to knowledge acquisition. This system was used as an interactive advisory tool with 680 rules.

However, traditional rule-based expert systems have some limitations when applied to the orthodontic diagnosis and treatment planning. Knowledge acquisition has become a major bottleneck in developing expert systems. This is due to the fact that the selection of knowledge (i.e., feature extraction) is difficult, the process for such extraction is time-consuming, and the rules were created based expertise knowledge through a trial-and-error approach. To resolve these issues, previous studies [12, 13] employed a new expert system concept called "case-based reasoning." This system uses a stored data bank of previously treated cases as a source of knowledge for solving new treatment problems. Case-based reasoning is based on the paradigm of human thought in cognitive psychology that contends that human experts derive their knowledge from solving numerous cases in their domain of focus. Although humans may generalize patterns of cases into rules, the principle unit of knowledge is the case. Thus, reasoning is made by analogy or association with the solutions to previous similar cases. Case-based reasoning can be particularly useful in areas where traditional rule-based reasoning is relatively weak, such as knowledge acquisition, machine learning, and reasoning with incomplete information.

The artificial neural network (ANN) is a system developed based on the basic model of the brain function in cognitive science in 1951 [14] with the goal of creating a "more generalized system" similar to the human brain. The ANN contains an algorithm for pattern recognition called "perceptron" [15, 16] (Fig. 6.2) and uses information received from the outside world rather than applying a set of formal decision rules; for example, a previously developed neural network was trained to make decisions regarding patients' lower third molar teeth by "informing" it of decisions that a consultant actually made [17]. The need for tooth extraction before orthodontic treatment was modeled using an ANN [18].

The issue to extract or not in orthodontics is often used as the topic of choice when devel-

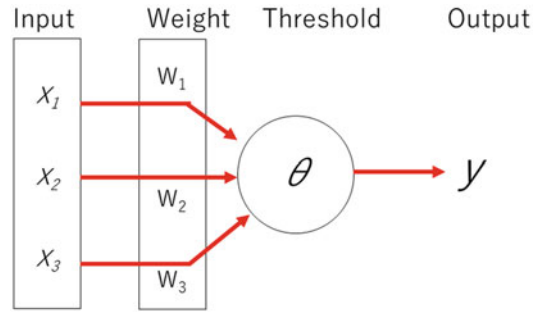


Fig. 6.2 Perceptron

oping an automatic treatment planning system. A previous study [19] developed a system to predict whether or not tooth extraction should be performed and, if extraction was chosen, to determine which tooth/teeth should be extracted. The authors resolved this issue as an optimization problem that maximized the objective function, defined as the rate of coincidence between the experts' answer and model's answer. They employed pattern matching of the feature vectors, representing the patient status, and the corresponding expertise answer was used as the system's output.

Recent works [20, 21] have employed multilayer perceptron artificial neural networks to predict orthodontic treatment plans. A previous study used only 156 patients for 12 cephalometric variables and additional 6 indexes, successfully identifying 5 categories of tooth extraction (nonextraction; extraction of #14, #24, #34, #44; that of #14, #24; that of #15, #25, #35, #45; and that of #14, #24, #35, #45) with 93% accuracy. Another study [20] found model success rates of 93% for the diagnosis of extraction vs. nonextraction and 84% for the detailed diagnosis of the extraction patterns, including the determination of extraction-nonextraction, extraction patterns, and anchorage patterns. These results indicate that there are clear patient patterns corresponding to treatment patterns, and it has become useful to store data necessary for problem-solving and process it on a computer to improve efficiency and convert it into knowledge that is easy for users to understand.

In medicine, most patient data have been described using natural languages, such as clinical findings and findings from images. In previous studies, feature extraction and data annotation were manually conducted by experts, which was time-consuming. It is therefore difficult to effectively use IT resources in medical fields in order to solve complex patient problems.

In the next chapter, we will present preliminary findings regarding the development of fully automated treatment planning for all orthodontic patients using NLP.

6.3 Overview of Our Automated Orthodontic Diagnosis System

As described in the Introduction, a certain regularity is apparent in the logic structure involved in making a medical diagnosis and planning treatment. We have been developing several AIs in order to fully automate orthodontic diagnoses.

A schematic illustration of such a system is shown in Fig. 6.3. The first step of this system is the creation of the medical findings from the diagnostic material. For example, an AI system that analyzes the patient's craniofacial characteristics was developed based on a system that automatically recognizes anatomical landmarks (please see Chap. 5). Distances and/or angles of lines connected to landmarks were measured and compared with sex- and age-matched controls, which is the same approach adopted by human orthodontists. Abnormal values are interpreted as linguistic descriptions. Another AI system recognizes facial photographs using CNNs and recurrent neural networks (RNNs) (please see Chap. 5) and creates facial descriptions as linguistic descriptions or classes. Linguistic descriptions of medical charts and patients' basic information are then collected. These descriptions, classes, and/or images are used to make the orthodontic diagnosis and develop a treatment plan.

In the next chapter, we will introduce the last part of the AI system: a system that automatically designs an orthodontic treatment plan based on a document of medical findings describing a patient's condition.

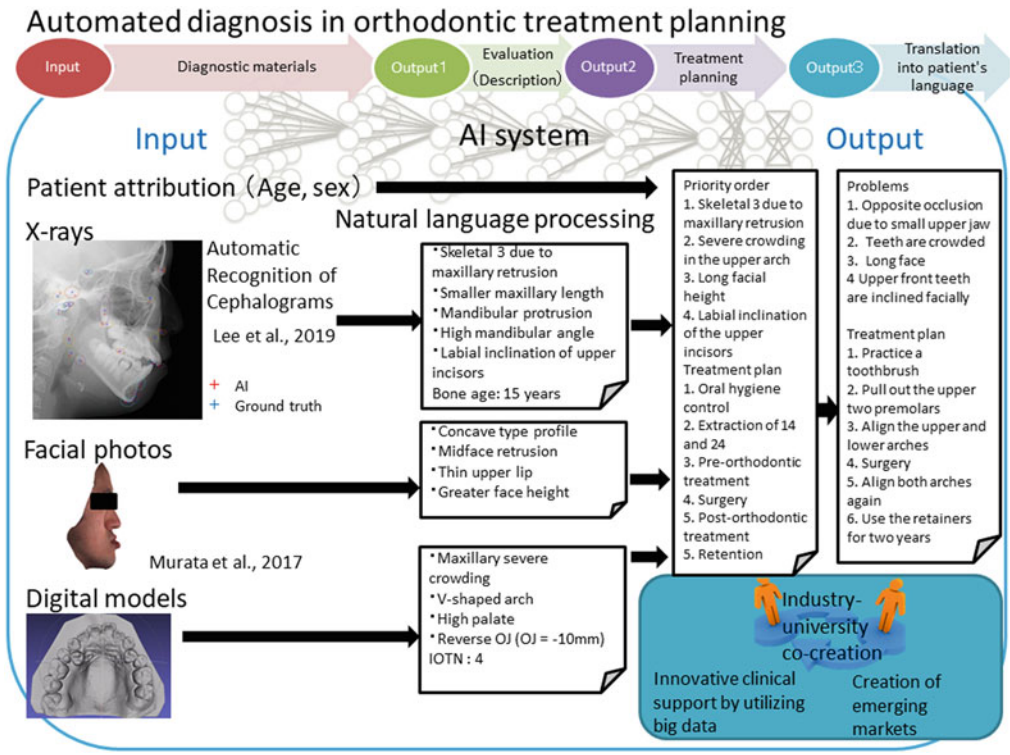


Fig. 6.3 Overview of our automated orthodontic diagnosis system

6.4 A System that Automatically Designs an Orthodontic Treatment Plan Based on a Document of Medical Findings Describing a Patient's Condition

6.4.1 Dataset and Problem Setting

In this study, we constructed a system that automatically designs an orthodontic treatment plan based on a document of medical findings describing a patient's condition. As the medical findings are documented in freewriting by the dentist, the feature vectors are extracted from the text using NLP, and a treatment plan is generated using machine learning models.

Text-to-text generation tasks in NLP, such as machine translation [22] and automatic summarization [23], require millions to tens of millions of sentence pairs to optimize deep learning models in an end-to-end manner. However, because it is difficult to collect pairs of medical findings and treatment plans on a large scale, only 1000 pairs were available for the present study. Therefore, we divided the task of generating a treatment plan from documented medical findings into the following two subtasks for more efficient resolution:

1. Medical Findings Summarization task: List patient orthodontic problems with priorities
2. Treatment Planning task: List treatment items according to the treatment procedure

In subtask 1, we listed the patient orthodontic problems and their priorities shown in Fig. 6.4b based on the medical findings shown in Fig. 6.4a. This subtask was considered to be an automatic summarization problem that generates a short document from a long document. However, as mentioned above, it is difficult to solve text-to-text generation problems using small datasets. Therefore, we arranged all of the orthodontic

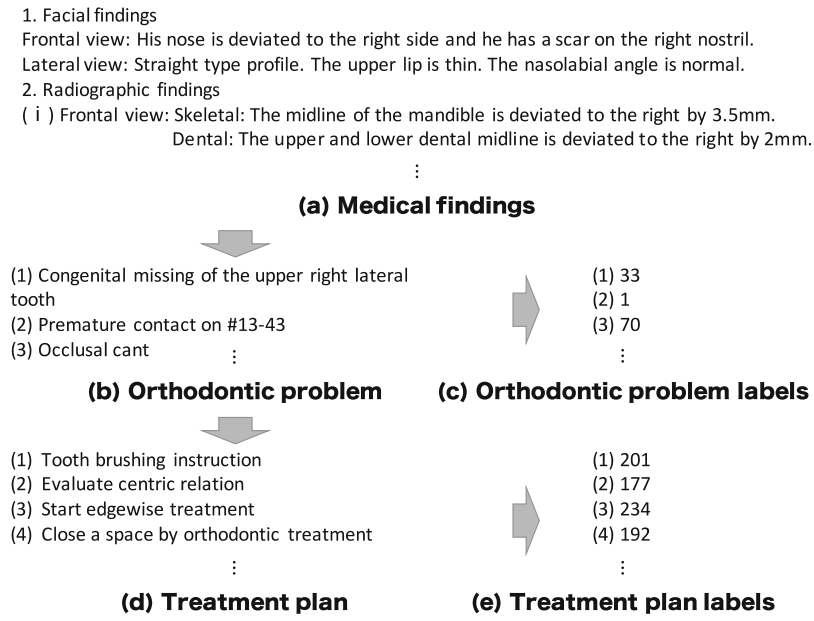


Fig. 6.4 Dataset

problems included in the dataset and classified them into the 300 orthodontic problem labels shown in Fig. 6.4c. Subtask 1 was then redefined as a document classification task for giving the text of the medical findings its appropriate corresponding orthodontic problem labels. Given that patients generally have multiple orthodontic problems, this study dealt with a multi-label document classification task.

In subtask 2, we listed the treatment items shown in Fig. 6.4d according to the treatment procedure from the summary of the medical findings shown in Fig. 6.4b. This subtask was considered to be a document-to-document machine translation problem. However, as with subtask 1, it was difficult to perform text-to-text generation using such a small dataset. Therefore, we arranged all of the orthodontic problems included in the dataset and classified them into the 300 orthodontic problem labels shown in Fig. 6.4e in a similar manner to subtask 1. Subtask 2 was then redefined as a sentence-to-sentence machine translation task that generated the sequence of the treatment items from the sequence of the orthodontic problem labels. Using the labels as the input and output instead of raw text can reduce both the amount of tokens and the length of the sequence.

6.4.2 Medical Findings Summarization Task

In this section, we consider the medical findings summarization task [24] as multi-label document classification that assigns the patient orthodontic problem labels shown in Fig. 6.4c based on the medical findings shown in Fig. 6.4a.

6.4.2.1 Methods

In document classification tasks, the input text is converted into a feature vector, and the class to which the text belongs is determined using a classifier based on a machine learning model. There are two main approaches to converting a document into a feature vector. One represents a document as a set of words, and the other represents a document as a set of sentences.

Word-Based Vector Representation of Documents

The most common approach to representing the document as a set of words is the bag-of-words approach. In this method, each word is first converted to a unique ID, and a feature vector is then constructed with the rule that if the i-th indexed word appears in the document, the i-th dimension of the feature vector value is set to 1; otherwise,

the i-th dimension's value is set to 0. This construction rule assumes that each word has equal importance in the document. To avoid the naïve word importance assumption in the bag-of-words approach, the term frequency-inverse document frequency (TFIDF) method also considers the importance of the words in the document.

$$\mathrm{TFIDF} = \frac{\mathrm{freq}\left(w_i, d_j\right)}{\sum_{w_k \in d_j} \mathrm{freq}\left(w_k, d_j\right)} \times \log \frac{N}{\mathrm{df}\left(w_i\right)}$$

Here, freq denotes the number of occurrences of word w in document d, df represents the number of documents in which word w appears, and N represents the number of all documents. The first term (TF) of the equation emphasizes the words that appear frequently in the target document. The second term (IDF) neglects the words that appear in many documents.

Recently, distributed word representations (word embeddings) that represent a word as a dense vector of hundreds of dimensions have been frequently used. In word2vec (w2v) [25], which is a typical approach to word embedding, the vector representation of words is learned by a neural network that estimates the context words from a word by following the distributional hypothesis [26]. The distributional hypothesis assumes that similar words should appear in similar contexts. In word embedding-based document classification [27], the document vector is constructed by averaging the vectors of each word appearing in the document for each dimension or taking the maximum value of each dimension.

Sentence-Based Vector Representation of Documents

Another approach represents a document as a set of sentences. This approach uses deep learning techniques, such as a RNNs, to construct a sentence embedding from word embeddings. Unsupervised learning methods, such as Skip-Thought [28], extend the distributional hypothesis in word embedding to the sentence level, learning vector representation of a sentence through the optimization of a neural network that estimates context sentences from a single sentence. Supervised learning methods, such as InferSent [29], learn vector representation of a sentence during the optimization of a neural network while predicting the semantic relationship between sentences. Multitask learning methods, such as the universal sentence encoder [30], enable computers to learn general-purpose sentence embeddings (representations), which are useful for various applications by training both unsupervised and supervised tasks with a single model.

Sentence embeddings are explained using Skip-Thought as an example. In Fig. 6.5, w2v (word2vec) converts a word into a word embedding. The RNN is a neural network that receives a word embedding as input and transfers information recursively from the beginning to the end of the sentence. <EOS> is a special token representing the end of the sentence, and the hidden layer of the RNN at the <EOS> time step can be used as the sentence representation.

In sentence embedding-based document classification, the document vector is constructed by averaging each dimension or taking the maximum

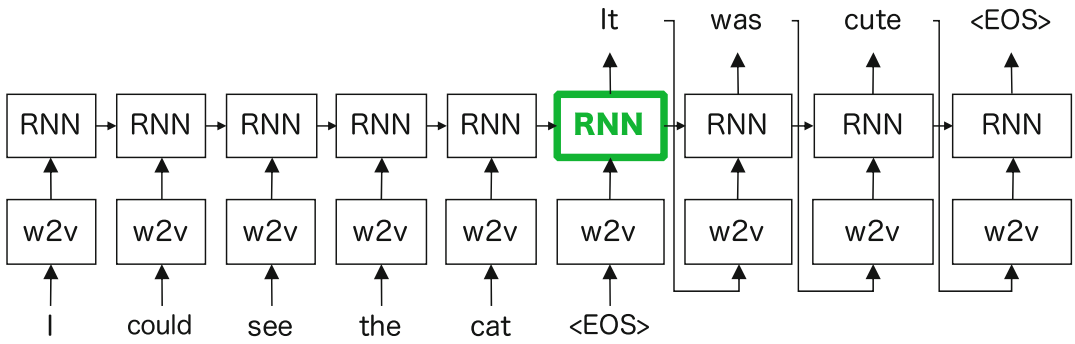


Fig. 6.5 *Recurrent* neural network for training sentence embedding

value for each dimension for the vectors of each sentence appearing in the document.

6.4.2.2 Experiments

The medical findings of about 1000 documents in Japanese were divided in an 8:1:1 ratio for training, development, and evaluation, respectively. MeCab [31] was used for word segmentation, and the support vector machine (SVM) [32] implemented in Scikit-learn [33] was used for the classifier. To apply the SVM system to the multi-label classification task, we used the binary relevance solution implemented in Scikit-multilearn [34].

By constructing the feature vectors via the bag-of-words approach using 2000 words with a high frequency in the training dataset, the orthodontic problem labels were classified with the following performance: precision = 0.667, recall = 0.311, and MicroF1 = 0.424. The bag-of-words method performed better than other methods, including sentence-based methods.

6.4.3 Treatment Planning Task

In this section, we address the treatment planning task as sequence-to-sequence learning, which generates a sequence of treatment item labels shown in Fig. 6.4e from a sequence of orthodontic problem labels shown in Fig. 6.4c.

6.4.3.1 Methods

In machine translation, a sequence-to-sequence model that maps a sequence of words in the source language to a sequence of words in the target language is trained. In this study, we considered a sequence of orthodontic problem labels as a sequence of words in the source language and a sequence of treatment item labels as a sequence of words in the target language, thereby generating a treatment plan using machine translation techniques. We use machine translation models based on RNNs [35] and self-attention networks (SANs) [22], which are common in deep learning-based machine translation.

In RNN-based machine translation, sequences are generated in the same manner as in Skip-Thought (Fig. 6.5). Word vectors are sequentially input into the RNN, and information is recursively transferred to construct a sentence vector of an input sentence. Words are then generated sequentially based on this sentence vector. With Skip-Thought, context sentences are generated, and with machine translation, a target language sentence is generated.

In contrast to RNNs, however, which model words sequentially from the beginning of the sentence (Fig. 6.6a), SAN-based machine translation models the entire sentence simultaneously with weight, as shown in Fig. 6.6b. In recent years, SANs have shown higher performance than RNNs for many NLP applications, although there remain tasks where RNNs are more effective, such as settings with low resources.

6.4.3.2 Experiments

As in Section 2.2, about 1000 sequence pairs of the dataset were divided in an 8:1:1 ratio for training, development, and evaluation, respectively. As both input and output use labels instead of raw text, word segmentation is not necessary for these experiments. The sequence-to-sequence models based on the RNN and SAN implemented in Sockeye [36] were used.

By generating a sequence of treatment item labels from a sequence of orthodontic problem labels using the RNN model, it was possible to automatically design a treatment plan with the following performance: precision = 0.435, recall = 0.430, and MicroF1 = 0.413. Similarly,

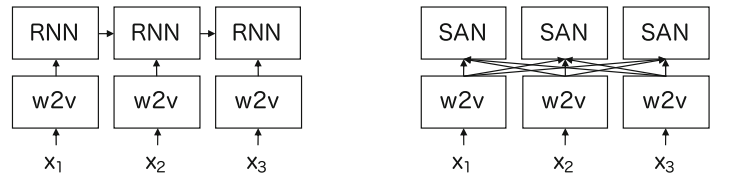


Fig. 6.6 Differences in sentence vector composition between RNN and SAN

the SAN model achieved the following performance: Precision = 0.436, Recall = 0.412, and MicroF1 = 0.412.

6.5 Conclusion

Automated treatment planning will reduce inter-planner variability and the planning time allocated for optimization to improve the plan quality. In this chapter, we described how to perform automated treatment planning. Automated treatment planning was divided into several AI systems, and we used NLP to integrate the assessments of different diagnostic materials. The merit of using NLP is that we can include clinical findings, which are usually presented as linguistic descriptions. By generating a sequence of treatment item labels from a sequence of orthodontic problem labels using the RNN model, it was possible to automatically design a treatment plan with a precision of 0.44 and recall of 0.43. This result indicates that this system would be effective for providing automated support for treatment planning of orthodontists. As treatment plans can vary even among human experts, we will compare the system's output with the human experts' answer in future publications.

References

1. Takada K. Elements of orthodontics. Medigit: Osaka; 2010.
2. Proffit WR. Contemporary orthodontics. 6th. ed. St. Louis, MO: Mosby; 2018.
3. Yu VL, Buchanan BG, Shortliffe EH, Wraith SM, Davis R, Scott AC, Cohen SN. Evaluating the performance of a computer-based consultant. Comput Programs Biomed. 1979;9(1):95–102.
4. Fischler MA, Firschein O. Intelligence: the eye, the brain and the computer: Addison-Wesley; 1987.
5. Sims-Williams JH, Brown ID, Matthewman A, Stephens CD. A computer controlled expert system for orthodontic advice. Br Dent J. 1987;163:161–6.
6. Zadeh LA. Fuzzy sets. Inf Control. 1965;8:338–53.
7. Brown ID, Adams SR, Stephens CD, Erritt SJ, Sims-Williams JH. The initial use of a computer controlled expert system in the treatment planning of class II division 1 malocclusion. Br J Orthod. 1991;18:1–7.
8. Stephens CD, Mackin N, Sims-Williams JH. The development and validation of an orthodontic expert system. Br J Orthod. 1996;23:1–9.
9. Takada K, Sorihashi Y, Akcam MO. Orthodontic treatment planning: its rationale for inference. In: Carels C, Willems G, editors. The future of orthodontics. Belgium: Leuven University Press; 1998. p. 203–21.
10. Akcam MO, Takada K. Fuzzy modelling for selecting headgear types. Eur J Orthod. 2002;24:99–106.
11. Poon KC, Freer TJ. EICO-1: an orthodontist-maintained expert system in clinical orthodontics. Aust Orthod J. 1999;15:219–28.
12. Hammond RM, Freer TJ. Application of a case-based expert system to orthodontic diagnosis and treatment planning: a review of the literature. Aust Orthod J. 1996;14:150–3.
13. Hammond RM, Freer TJ. Application of a case-based expert system to orthodontic diagnosis and treatment planning. Aust Orthod J. 1997;14:229–34.
14. Minsky M. A neural-analogue calculator based upon a probability model of reinforcement. Cambridge, MA: Psychological Laboratories, Harvard University; 1952.
15. Rosenblatt F. Principles of Neurodynamics. Washington, D.C.: Spartan Books; 1961.
16. Minsky M, Papert S. Perceptrons–an introduction to computational geometry. Cambridge, MA: The MIT Press; 1969.
17. Brickley MR, Shepherd JP, Armstrong RA. Neural networks: a new technique for development of decision support systems in dentistry. J Dent. 1998;26:305–9.
18. Xie X, Wang L, Wang A. Artificial neural network modeling for deciding if extractions are necessary prior to orthodontic treatment angle orthod. 2010;80:262–266.
19. Takada K, Yagi M, Horiguchi E. Computational formulation of orthodontic tooth-extraction decisions. Angle Orthod. 2009;79(5):885–91.
20. Li P, Kong D, Tang T, Su D, Yang P, Wang H, Zhao Z, Liu Y. Orthodontic treatment planning based on artificial neural networks. Sci Rep. 2019;9:2037.
21. Junga SK, Kimb TW. New approach for the diagnosis of extractions with neural network machine learning. Am J Orthod Dentfac Orthop. 149(1):127–33.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł. Attention is all you need. Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2017:5998–6008.
23. Rush AM, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2015:379–8.
24. Kajiwara T, Tanikawa C, Shimizu Y, Chu C, Yamashiro T, Nagahara H. Using natural language pro-

cessing to develop an automated orthodontic diagnostic systems. arXiv:1905.13601. 2019;1–6.
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations (ICLR). 2013;1–12.
26. Harris ZS. Distributional structure. Word. 1954;10:146–62.
27. Shen D, Wang G, Wang W, Min MR, Su Q, Zhang Y, Li C, Henao R, Carin L. Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). 2018;440–450.
28. Kiros R, Zhu Y, Salakhutdinov R, Zemel RS, Torralba A, Urtasun R, Fidler S. Skip-thought vectors. Proceedings of the Advances in Neural Information Processing Systems (NIPS). 2015;3294–3302.
29. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised learning of universal sentence representations from natural language inference data. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2017;670–680.
30. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Strope B, Kurzweil R. Universal sentence encoder for English. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018;169–174.
31. Kudo T, Yamamoto K, Matsumoto Y. Applying conditional random fields to Japanese morphological analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2004:230–7.
32. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the Annual Workshop on Computational Learning Theory. 1992:144–52.
33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
34. Szymański P, Kajdanowicz T. Scikit-multilearn: a Scikit-based python environment for performing multi-label classification. J Mach Learn Res. 2019;20:1–22.
35. Bahdanau D, Bengio CK. Y. Neural machine translation by jointly learning to align and translate. Proceedings of the International Conference on Learning Representations (ICLR). 2015:1–15.
36. Hieber F, Domhan T, Denkowski M, Vilar D, Sokolov A, Clifton A, Post M. Sockeye: a toolkit for neural machine translation. arXiv:1712.05690. 2017;1–18.

Machine Learning in Orthodontics: A New Approach to the Extraction Decision

7

Mary Lanier Zaytoun Berne, Feng-Chang Lin, Yi Li, Tai-Hsien Wu, Esther Chien, and Ching-Chang Ko

7.1 Introduction

The extraction of permanent teeth versus expansion of the dental arch to correct dental malalignment has long been a source of debate in the orthodontic community. The origins of this debate can be traced back to the early twentieth century when the profession of orthodontics had just begun to be seen as a science. Edward Angle, "the father of modern orthodontics," had a vision for orthodontic treatment centered on aligning all 32 teeth in perfect Class I occlusion. His treatment philosophy was predicated on the principles that "balance, harmony, and proportions of the mouth require that there shall be a full complement of teeth." He advocated for a non-extraction approach stating that with tooth movement, the jaw and alveolar bone would grow to compensate and the corresponding tissues would also adapt to their new positions.

However, Angle's contemporary, Calvin Case, opposed this orthodontic treatment philosophy and defended extractions as the ideal treatment to correct facial deformities and address malalignment, especially in cases of dental protrusion. Corroborating Cases' viewpoint, Charles Tweed initially chose to follow Angle's treatment philosophy of arch expansion, but later in his career realized that many of these patients were experiencing significant posttreatment relapse. Tweed retreated a number of these relapse patients by extracting four premolars and obtained a satisfactory and more stable occlusal result. Therefore, Case and Tweed were among many orthodontists of this era who advocated for extractions as a means to achieve adequate stability. This philosophy became commonplace, ultimately reaching its peak between the 1950s and 1960s when approximately 50 percent of orthodontic patients were treated with extractions, usually of the first premolars. In fact, a 40-year study by Dr. Proffit

M. L. Z. Berne
Department of Orthodontics, Adam School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

F.-C. Lin (✉)
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: flin33@email.unc.edu

Y. Li
Department of Biostatistics and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

T.-H. Wu · E. Chien · C.-C. Ko
Division of Orthodontics, College of Dentistry, The Ohio State University, Columbus, OH, USA

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_7

analyzed the fluctuation of extraction frequencies at the University of North Carolina. Proffit found that the number of patients with extraction of all four first premolars increased from 10% in 1953 to 50% in 1963, remained at 35–45% until the early 1980s, then declined sharply back to the 1950s level by 1993 [1]. The increase in first premolar extractions occurred primarily as part of a search for greater long-term stability, while the recent decline seems to be due to a number of factors. Greater concern about the impact of extraction on facial esthetics, data to suggest that extraction does not guarantee stability, concern about temporomandibular dysfunction (TMD), and changes in technique all appear to have played a role. With appropriate orthodontic mechanics, many patients with Class I crowding can be treated satisfactorily with or without premolar extraction [2].

Later, as the profession continued to advance with the influx of new technologies and techniques, treatment modalities improved. Clinicians began to understand that soft tissue balance was critical for successful orthodontic treatment; ideal results were no longer predicated on dental occlusion alone. Dr. Ackerman, Dr. Sarver, and Dr. Proffit touted this monumental change in orthodontic thought as "the soft tissue paradigm" [3]. They surmise that soft tissue, not hard tissue skeletal discrepancies, is the major limitation to orthodontic treatment and "therapeutic modifiability." Soft tissue relationships and functions (such as pressure exerted by the lips, cheeks, and tongue, the contours of the soft tissue, especially the lips and the nose, and anterior tooth display during facial animation) as well as soft tissue adaptions (from an equilibrium, periodontal, TMD, facial balance, and anterior tooth display standpoint) are more limiting than the anatomic or hard tissue restrictions to treatment. With this new outlook, as well as emerging data that suggested that extraction does not guarantee stability, extraction rates in orthodontics began to fall.

This retrospective look on the history of orthodontics clearly demonstrates that the decision for or against extraction is not a simple one. It is evident that the rates of and reasons for tooth extraction versus arch expansion have fluctuated significantly over the years. Distilling down the philosophic beliefs of these historical debates, orthodontic treatment planning today rests on three guiding principles: occlusion, stability, and soft tissue balance. In taking each of these three principles equally into account, the orthodontic community today understands that some patients will benefit from extraction treatment while others will not.

But determining who those patients are and how they fall into each category is still highly debated. When deciding whether or not to extract permanent teeth, orthodontists must weigh many factors including spacing or crowding, overjet, overbite, occlusal stability, temporomandibular dysfunction, periodontal health, facial esthetics, smile arc, and systemic health among others. One of the factors that make extractions one of the most difficult treatments to diagnose is the fact that a significant proportion of orthodontic cases fit in the "borderline" category—meaning both extraction and non-extraction treatment plans can be considered. Orthodontists have learned to evaluate all of the predisposing clinical factors and make an empirical treatment decision by relying on their own clinical experiences, their training, and their personal treatment philosophies. And orthodontists tend to have a strong preference for, or even bias toward, their chosen treatment modality of extraction versus non-extraction. But historically there has been very little research to quantify this decision and determine on a case-by-case basis, from a data-driven standpoint, which treatment modality is the best fit for each individual patient. In the age of evidence-based science and personalized medicine, the patients and the profession demand that this decision-making process be improved upon. The extraction versus expansion decision needs to be made by eliminating subjectivity and personal bias in favor of an empirical, statistically based method for aiding in this decision.

To this end, Dr. Guez conducted a follow-up to Dr. Proffit's study, seeking to identify the main factors that lead to orthodontic extractions. Guez et al. collected the data of orthodontic patients treated at the University of North Carolina from

the years 2000 to 2011 [4]. Among this patient data set, they found that the factors leading to statistically significant odds of extraction were being of African-American ethnicity (as compared to the Caucasian reference group) as well as Angle's classification (skeletal and dental anteroposterior), initial overjet, overbite, and amount of maxillary and mandibular crowding. Their findings regarding the rates of extraction during this time period were as follows: there was a continued mild decrease in orthodontic extractions, both in overall extractions (leveling off near 25% after 2006) and in four premolar extractions (just above 10%).

7.2 Logistic Regression

We continued the research of Guez and collaborators by identifying thresholds (cutoffs) for these predisposing factors: overbite, overjet, maxillary crowding, and mandibular crowding. In this study, the aggregate data of 2003 consecutively treated orthodontic patients in the Graduate Clinic of the Orthodontic Department at the University of North Carolina from the years 2000 to 2011 were analyzed. For each patient treated in this clinic, pretreatment variables were recorded and stored in a digital database. These records, which include demographic information, patient interview answers, and clinical examination findings, are standardized and stored in a secured digital database. Among the clinical examination findings are factors that influence extraction rates. The outcome was the decision to extract teeth, or alternatively, to not extract teeth, for orthodontic purposes.

A receiver operating characteristic (ROC) analysis was used to evaluate the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for the individual clinical variables of overjet, overbite, maxillary crowding, and mandibular crowding. ROC analysis operates under the principle that as sensitivity (the ability to detect true positive) is increased, specificity (the ability to detect true negative) decreases. Therefore, in a series of clinically relevant values for each variable, sensitivity and specificity were calculated and ROC analysis was used to determine the optimal cutoff (threshold) of each variable by minimizing the difference between these two characteristics. Using an accrued score from the four dichotomized clinical variables (scoring one point for each observed value over the threshold), we developed a decision tree model that can be used to determine whether or not to extract teeth. In addition, we developed a logistic regression formula for the calculation of the probability of tooth extraction using measured clinical values (i.e., amount of overjet in mm) of each variable.

By maximizing the difference between sensitivity and specificity as described, the cutoffs were determined to be 4.5 mm for overjet, 3.5 mm for overbite, 6.5 mm for maxillary crowding, and 5.5 mm for mandibular crowding (Tables 7.1, 7.2, 7.3, 7.4).

We also built a dichotomous tree and a logistic regression formula that combined these factors into chairside clinical aids for making these decisions (Figs. 7.1 and 7.2). For the decision tree, each value that met or exceeded the cutoff for that factor was assigned a score of 1 and the composite score was compiled by calculating the sum of the

Table 7.1 Overjet values

Overjet (mm)	Sensitivity	1—Specificity	−log(Sensitivity* Specificity)
1.50	0.895	0.915	1.119
2.50	0.757	0.747	0.718
3.50	0.621	0.500	0.508
4.50	**0.460**	**0.326**	**0.508**
5.50	0.304	0.194	0.611

A cutoff of 4.5 mm yields a sensitivity of 46% and a specificity of 67%

Table 7.2 Overbite values

Overbite (mm)	Sensitivity	1—Specificity	−log(Sensitivity* Specificity)
1.50	0.227	0.177	0.704
2.50	0.394	0.289	0.553
3.50	**0.606**	**0.504**	**0.522**
4.50	0.768	0.703	0.642
5.50	0.878	0.851	0.884

A cutoff of 3.5 mm yields a sensitivity of 61% and a specificity of 50%

Composite Score Calculation	
Overjet	> 4.5mm
Overbite	< 3.5mm
Maxillary crowding	> 6.5mm
Mandibular crowding	> 5.5mm

* for each clinical factor scoring >/< the indicated value, a score of 1 is assigned. The sum of the scores for each clinical factor = the composite score.

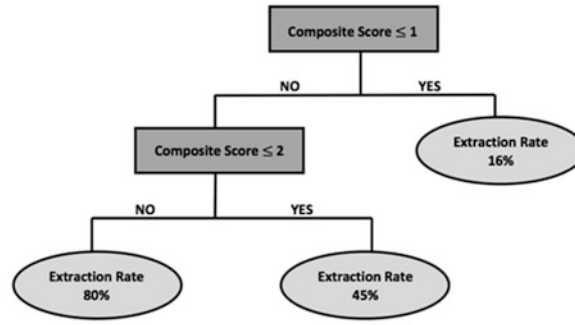


Fig. 7.1 Composite score decision tree model. One point is scored for each outcome and the sum is the composite score for that patient

Table 7.3 Mandibular crowding values

Mandibular crowding (mm)	Sensitivity	1—Specificity	−log (Sensitivity* Specificity)
3.50	0.734	0.442	0.893
4.50	0.615	0.271	0.802
5.50	**0.486**	**0.138**	**0.869**
6.50	0.347	0.077	1.138
7.50	0.216	0.034	1.566

A cutoff of 5.5 mm yields a sensitivity of 49% and a specificity of 86%

Table 7.4 Maxillary crowding values

Maxillary crowding (mm)	Sensitivity	1—Specificity	−log (Sensitivity* Specificity)
3.50	0.705	0.473	0.990
4.50	0.603	0.315	0.883
5.50	0.497	0.166	0.881
6.50	**0.397**	**0.091**	**1.019**
7.50	0.277	0.057	1.344

A cutoff of 6.5 mm yields a sensitivity of 40% and a specificity of 91%

factors. A low chance of extraction (16%) was indicated when the composite score was less than 1 while a high chance of extraction (80%) was predicted when the composite score was greater than 2.

For the logistic regression formula, in cross tabulating extractions determined by the formula and the actual extractions performed in our patient sample, the extraction cutoff was determined to be 36% (Table 7.5).

$$\log\left(\frac{p}{1-p}\right) = -1.6 + 0.2(overjet\ in\ mm) - 0.2\ (overbite\ in\ mm) + 0.1\ (max\ crowding\ in\ mm) + 0.1\ (mand\ crowding\ in\ mm)$$

Fig. 7.2 Logistic regression formula for probability of extractions

Table 7.5 Extraction by clinical formula x Actual extraction cross-tabulation

		Actual extractions	
		No	Yes
Extraction by logistic regression formula	No	121886%	24847%
	Yes	20614%	27853%

The optimal cutoff probability for extraction decisions is 36% (53% sensitivity and 86% specificity)

In order to demonstrate this method's potential clinical usefulness and viability, patients that were in treatment in the University of North Carolina Graduate Orthodontic Clinic were selected as examples. The patients selected represented borderline cases, in which practitioners typically differ in their propensity for extractions. One such example showing the utility of these equations is shown below.

Patient A.R., a 15-year-old Hispanic male, presented to the orthodontic clinic with moderate to severe crowding and a chief complaint of "I don't like the appearance of my teeth. The bottom ones are crooked and the top canines stick out." He had skeletal and dental Class I relationships with 5 mm overjet and 4 mm overbite. His upper left first premolar was in a Brodie bite, and

Fig. 7.3 Patient A.R. intraoral photos

his mandibular midline mildly deviated from the midsagittal plane (Fig. 7.3).

A.R. is an example of a borderline extraction case because of his ectopically erupting maxillary canines and moderate-to-severe mandibular crowding.

Ultimately, due to the amount of crowding present in the maxillary and mandibular arches, the decision was made to extract maxillary and mandibular first premolars. Treatment included full fixed appliances using self-ligating brackets with an 0.022 slot and sliding mechanics.

Using our evidence-based model, patient A.R.'s composite score was calculated to be a 2. Given the composite score decision tree, there was a 45% chance for extraction in this case. The logistic regression formula also gives the extraction probability p of 45%, which indicated extraction. In this case, our clinical decision was confirmed by our predictive models (Fig. 7.4).

This case represents one example showing why we believe these models have some clinical utility. However, as discussed previously, we recognize that the extraction decision is very complex, and the four factors analyzed are not the only variables that influence this decision. Clinicians frequently cite other factors that weigh heavily in their decision-making, such as smile esthetics and gingival display, periodontal health, TMJ function, dental and lip protrusion, age, gender, and ethnicity, among many other factors [5, 6]. In order to continue to advance this research to a level providing clinical relevance, the next step is to incorporate machine learning and big data, which will allow us to incorporate hundreds of variables at a time. A machine learning model would represent a huge advancement and improvement upon the dichotomous tree and logistic regression models previously mentioned and increase the clinical efficacy of this research.

7.3 Machine Learning

First, let us discuss what exactly is machine learning, and what machine learning methods can be utilized for these ends. Machine learning is a branch of artificial intelligence that employs a variety of statistical and optimization techniques that allow computers to "learn" from past examples and to detect hard-to-discern patterns and relationships from complex data sets. Machine learning is a rapidly growing field, with applications in every facet of our lives. It is used in other areas of dentistry such as for the diagnosis of bisphosphonate-related osteonecrosis of the jaw and for the detection of caries [7, 8]. Machine learning has also been utilized in the medical world, and some examples of such uses are the diagnosis of cancer based on imaging of tumor biopsies, the detection of potential adverse drug reactions, assessing risk prediction for cardiovascular disease, among many others [9–11]. Just as

Overjet (mm)	Overbite (mm)	Maxillary Crowding (mm)	Mandibular Crowding (mm)	Composite Score	Extraction Probability (logistic regression formula)	Model Prediction	Clinical Decision
5	4	6	6	2	45%	Extract	Extract U/L4s

$$log\left(\frac{p}{1-p}\right) = -1.6 + 0.2\,(5mm) - 0.2\,(4mm) + 0.1\,(6mm) + 0.1\,(6mm)$$

$$log\left(\frac{p}{1-p}\right) = -0.20$$

$$p = 45\%$$

Composite Score Calculation		Score 1 or 0*
Overjet	> 4.5mm	1
Overbite	< 3.5mm	0
Maxillary Crowding	> 6.5mm	0
Mandibular Crowding	> 5.5mm	1
	SUM	2

*1 indicates threshold is met or exceeded for that factor.
0 indicates threshold is not met or exceeded for that factor.

Fig. 7.4 Patient A.R., clinical data, composite score, and logistic regression outcomes

these algorithms have been utilized for diagnosis in other areas of dentistry and medicine, we intend to utilize machine learning and big data to diagnose the extraction versus non-extraction decision in orthodontics.

7.3.1 Three-Layer Artificial Neural Network

In fact, there is precedent for using machine learning to approach the extraction decision. A previous study by Jung and Kim constructed four three-layer (an input layer, a hidden layer, and an output layer) neural networking models for this decision using the data of 156 patients, treated by one clinician, with input data consisting of 12 variables from cephalometric analysis and 6 variables from clinical findings [12]. Jung et al.'s models included (1) a model for determining whether or not to extract, (2) a model for determining differential extractions (defined as the extraction of all first premolars or all second premolars, "non-differential" vs. the extraction of upper first premolars and lower second premolars, "differential"), (3) a model for determining "more retraction" (defined as the extraction of first premolars instead of second premolars) in non-differential extraction cases, and (4) a model for determining "more retraction" in a differential extraction case (defined as the extraction of only upper premolars vs. the extraction of upper first premolars and lower second premolars). After three stages of model training, in which the data was divided into a "training set (64 patients)" and a "validation set (32 patients)", the decision-making success rate was calculated for each model. Model 1, extraction vs. non-extraction, demonstrated the highest decision-making accuracy at 93% for the test dataset.

While Jung introduced the concept of using machine learning for the orthodontic extraction decision, his models had many weaknesses. For

example, their models were overfitting at the end of their training progress, which implies that their dataset is too small. It is difficult to determine whether some of their models (e.g., their fourth classifier) were trained well even though they used an early stopping technique. Because of the small dataset size, only 12 variables were used as input data. However, 12 variables might not comprehensively express the patient's situation. Recently, Li et al. used a similar artificial neural network architecture with more advanced techniques such as incorporating a dropout layer to prevent overfitting [13]. In their study, their model has 24 variables including demographic, cephalometric, and soft tissue data, resulting in an accuracy of 94% based on 302 patients for the extraction vs. non-extraction decision. We seek to improve on these data by using a larger and more varied patient pool (almost 850 patients with treatment provided by 25 different clinicians, each with their own treatment modalities and philosophies) and more diversity in the input data (an average of 100 additional input variables in each of our models).

7.3.2 Comprehensive Survey

Before delving into the specifics of our study, let us acknowledge that there are countless machine learning approaches and algorithms, each with their own various strengths and weaknesses. For the purposes of this study, and after extensive modeling tests with various algorithms, we narrowed our focus to include two algorithms: Random Forest and Multilayer Perceptron (MLP). A third algorithm, Classical and Regression Trees (CART), used in the previous four-parameter study, was included as a reference method. Let us begin by explaining the function of each of these algorithms.

7.3.2.1 Classical and Regression Trees

CART is a nonparametric modeling technique for regression and classification problems that uses a decision tree. The decision tree makes sequential, hierarchical decisions using multiple predictors to arrive at a prediction on the outcome variable. It uses a set of consecutive binary rules to divide the sample into several subsamples. There are different algorithms that can be used to determine the best split variable at each node to divide the parent node into two child nodes. The selection of the split variable and split point can be chosen either simultaneously or one step after another. When there is no significant split variable nor split point for further splitting, the split is stopped. Consecutive splitting eventually grows the whole tree. Stopping and so-called pruning prevents the decision tree from overfitting. Figure 7.1 in the four-parameter study is derived from such an algorithm.

7.3.2.2 Random Forest

Random forest is a machine learning algorithm composed of a combination of random sampling and multiple decision trees. As described above, in each decision tree, a set of consecutive binary rules is established to divide the sample into several subsamples. Different from CART, each tree is trained independently with a randomly selected subset of the training sample with a subset of predictors. The tree is allowed to grow to the fullest diversity without pruning. The final classification, or output, is determined by taking the most common predictions from the terminal ensemble of each tree [14, 15].

7.3.2.3 Multilayer Perceptron

Multilayer Perceptron (MLP) is an artificial neural network algorithm patterned after the operation of neurons in the human brain. It can be used for classifying data and regression problems. The neural network models used in previous studies by Jung and Kim and by Li et al. are examples of MLP. A typical MLP is composed of an input layer, an output layer, and at least one hidden layer. The hidden layers typically consist of several neurons. The neural network receives an input and transforms it through a series of hidden layers. Except for the input layer, all the neurons in a layer are fully connected to all the neurons in the previous layer. Unlike a linear perceptron, MLP utilizes nonlinear activation in all neurons. The output layer uses certain activation functions (e.g., softmax function) to calculate a probability

for each class, which directly leads to the final classification.

7.3.2.4 Methods

Given this background, our study was set up in this way: The aggregate data of 842 consecutively treated, orthodontic patients in the graduate orthodontic clinic at the University of North Carolina from 2010 to 2013 were analyzed. Our patient population characteristics as well as many of the characteristics that are relevant in the extraction decision are listed in Table 7.6.

For each patient treated in this clinic, pretreatment variables—including demographic information, patient interview answers, and clinical examination findings—were recorded, standardized, and stored in a digital database. Additionally, cephalometric analysis was performed on initial cephalometric radiographs for each of these patients in Dolphin Imaging (Patterson Dental, St, Paul, Mn). Among the data gathered (demographics, clinical findings, and cephalometric analysis) are factors that influence extraction rates.

Four models were built utilizing the aforementioned machine learning algorithms. Each model incorporates a different machine learning algorithm and/or a different set of variables.

Model 1 is a CART using four measurable clinical variables that were deemed significant to the extraction decision in our previous study. Those input variables are overjet, overbite, maxillary crowding, and mandibular crowding.

Model 2 is a CART incorporating eleven variables. The input variables for this model include the four variables previously mentioned in Model 1 (overjet, overbite, maxillary crowding, and mandibular crowding) as well as additional clinical data (gingival attachment, curve of spee, skeletal anteroposterior relationship, and Angle classification) and demographic data (sex, race, and age).

Model 3 is a random forest algorithm with 117 input variables. The input variables for this model include the variables in Model 2 as well as the addition of 102 cephalometric variables from the initial cephalometric radiograph tracing.

Model 4 is a MLP algorithm with 117 input variables. This model uses the same variables as Model 3 but applies a different machine algorithm. In this study, the MLP algorithm is similar to the neural network model in Li et al.'s study discussed earlier [13]. Our model consists of a 117-neuron input layer, a 10-neuron hidden layer, a dropout layer with 0.5 probability, and a 2-neuron output layer. A batch normalization is used between the hidden layer and its activation function. All activation functions are hyperbolic (tanh), except a softmax function is used in the output layer. The loss function is binary cross-entropy.

In order to calculate average accuracy and to avoid overfitting the models, 10-fold cross-validation was performed. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal-sized subsamples. Of the 10 subsamples, 9 of them are used to train the model and the trained model is applied to the 10th subsample, the testing sample, to predict outcomes. This is repeated 10 times, with each of the 10 subsamples used exactly once as the validation (or testing) data. The average accuracy resulting from each of these testing sets is determined to be the model accuracy. By continually tuning hyper-parameters in this cross-validation, the model can be generalized to an unseen data set with more robust accuracy results.

ROC analysis was also used to evaluate the trade-off between sensitivity and specificity for each model. A high sensitivity (true positive) indicates the ability of the model to identify the patients who received extraction treatment, while a high specificity (true negative) indicates the ability of the model to identify the patients who received non-extraction treatment. Ideally, the best model would have both high sensitivity and high specificity.

7.3.2.5 Results and Discussion

In order to understand the true utility of these models in a clinical setting, the ability of each model to correctly predict the treatment modality for both extracted and non-extracted cases, referred to as model "balanced accuracy," defined as the arithmetic mean of sensitivity and specificity, was determined. The balanced accuracy of each model was calculated through 10-fold

Table 7.6 Participant characteristics

Variables in Model 2	Variables in Model 1	Characteristics	Overall median (IQR) or % ($n = 842$)	Non-extraction median (IQR) % ($n = 634$)	Extraction median (IQR) % ($n = 208$)	p-value*
Model 2	Model 1	Initial maxillary crowding (mm), median (IQR)	2 (−2 to 5)	2 (−2 to 4)	4 (1 to 7)	<0.001
		Initial mandibular crowding (mm)	3 (1 to 5)	3 (0 to 4)	5 (3 to 7)	<0.001
		Initial overbite (mm)	3 (2 to 5)	4 (2 to 5)	3 (2 to 4)	<0.001
		Initial overjet (mm)	4 (2 to 5)	3 (2 to 5)	4 (2 to 5)	0.017
		Sex (%)				
		Female	59.1	57.9	63.0	0.224
		Male	40.9	42.1	37.0	
		Race (%)				
		White	61.5	67.2	44.2	<0.001
		African American	15.7	14.8	18.3	
		Other	22.8	18.0	37.5	
		Age at start of treatment (y)	14.3 (12.9 to 17.1)	14.2 (12.9 to 17.1)	14.6 (13.1 to 16.7)	0.490
		Skeletal AP relationship (%)				
		Class I	53.7	56.0	46.6	0.005
		Class II	33.6	30.6	42.8	
		Class III	12.7	13.4	10.6	
		Angle Classification (%)				
		Class I	39.8	42.6	31.2	0.010
		Class II	51.1	49.2	56.7	
		Class III	9.1	8.2	12.0	
		Initial Curve of Spee (mm)	2 (2 to 3)	2 (1.3 to 3)	2 (2 to 3)	0.887
		Reduced Attached Gingiva (%)	17.1	16.4	19.2	0.405

Model 1 includes 4 variables (initial maxillary crowding, initial mandibular crowding, initial overbite, initial overjet). Model 2 includes all variables reported in this table. *Level of statistical significance set to $p = 0.05$

Table 7.7 Model types with accuracy, sensitivity, and specificity

	Model types	# of Variables included	Balanced accuracy	Sensitivity	Specificity
Model 1	CART	4	71.5%	60%	83%
Model 2	CART	11	66.5%	40%	93%
Model 3	Random Forest	117	72.5%	71%	74%
Model 4	MLP	117	71.5%	71%	72%

cross-validation. Model 1, CART with 4 variables, had a balanced accuracy of 71.5%; Model 2, CART with 11 variables, had a balanced accuracy of 66.5%; Model 3, Random Forest with 117 variables, had the highest balanced accuracy of 72.5%; and Model 4, MLP with 117 variables, had a balanced accuracy of 71.5%. The associated rates of sensitivity and specificity are seen in Table 7.7.

In addition to accuracy, identification of the most important variables for making the extraction decision can be made in the random forest algorithm. Variables that reduce the most errors when splitting were considered critical ones with high importance scores. Therefore, we were able to determine which variables the algorithm weighted most highly when making its extraction vs. non-extraction decision in Model 3. As a result, these variables were, in ranked order: (1) maxillary crowding, (2) mandibular crowding, (3) SN-GoGn value, (4) SN-GoGn deviation, and (5) maxillary unit length.

The first two variables, maxillary crowding and mandibular crowding, were standard clinical measurements, recorded in millimeters. The next two variables, SN-GoGn value and deviation, are measures gathered from each initial cephalometric tracing. SN-GoGn is an angular measurement of the skeletal vertical position. It is formed by intersecting the Gonion-Gnathion plane with the Sella-Nasion line, and it reflects the degree of inclination of the mandible relative to the anterior cranial base. A small angular measurement indicates a low plane angle, or "short" face. A larger angular measurement indicates a high mandibular plane angle, associated with a "long" face. "SN-GoGn deviation" specifically refers to the degree deviation of that particular measurement from the identified cephalometric norms, as defined by previous extensive research on the topic.

Based on these results, random forest (Model 3) appears to be the best model, showing the most promise for being a clinically utilized tool, for the extraction vs. non-extraction decision. Random forest had the highest balanced accuracy at 72.5%, with sensitivity and specificity values that remained relatively high as well at 71% and 74%, respectively. However, it is worth noting that a balanced accuracy of 72.5% still presents a barrier to viable implementation of this model in clinical decision making. As we saw before, the "borderline" cases are the most difficult cases to make a decision for extraction vs. non-extraction treatment. It is likely that at 72.5% accuracy, the cases that are not being accurately identified are those for which clinicians most desire empirical data—the borderline cases.

The rank of variable importance in the random forest algorithm of Model 3 empirically corroborates common clinical thinking that crowding, both maxillary and mandibular, is the most influential factor when deciding whether or not to extract teeth. It is interesting to note that the third and fourth most important variables, SN-GoGn value and deviation, are measures of the skeletal vertical position. This would indicate that the skeletal vertical, especially the extreme dolichofacial or brachyfacial presentations, plays a critical role in the algorithm's decision to predict extractions. This is a noteworthy finding because the previous research cited in this study did not include skeletal vertical considerations in their most prevalent factors that were considered in this decision. The results of this algorithm's variable importance would suggest that practitioners should weigh skeletalvertical position

even higher than anterior-posterior considerations when deciding whether or not to extract teeth.

One of the strengths of the models we used is the fact that they were created from a pool of more than 800 patients treated by approximately 25 different attending orthodontists using a variety of treatment modalities and varying treatment philosophies. Additionally, the demographics of our patient population are representative of the overall US population according to the 2010 US census. Both of these factors suggest that these equations could be applied accurately and reliably to the patient pool in average US orthodontic practices and to clinical orthodontic practices at large.

In terms of the strength of the input variables, these models, particularly Models 3 and 4, are comprehensive in their breadth of data and information. They incorporate nearly every clinical and demographic factor that a clinician must consider when making their decision for or against extraction. Incorporating all standard measurements of cephalometric tracing, including cephalometric soft tissue measurements, is a novel advance for machine learning algorithms in this field.

We must remember, machine learning is a relatively new and still growing field. As machine algorithms continue to develop and become more complex, these models, particularly deep learning, will continue to improve. As these models continue to advance, to further strengthen and increase their clinical relevance, the next step of this research would be to divide the patient pool into subgroups based on categories such as race, Angle classification, and "borderline cases." In doing this, we would be able to determine the model's accuracy in each category—for example, these models may have significantly increased accuracy in predicting extraction vs. non-extraction treatment in Angle Class II cases. If this example were found to be true, this finding could have major clinical implications in the way clinicians are able to utilize these models in the treatment of Class II cases.

7.4 Conclusion

As stated previously, the extraction vs. non-extraction decision is one of the most difficult clinical decisions that orthodontists face on a daily basis. We put forth that machine learning algorithms, particularly random forest, show promise in advancing the clinical decision-making process from a subjective decision to an evidence-based decision. This study advanced the research in this field by increasing the breadth of evidence available. To date, it has the largest patient pool and the most clinical variables on this subject, and four different models were tested. As algorithms continue to be further developed, we believe that the clinical utility of these models can change the landscape of orthodontic treatment planning.

References

1. Proffit WR. Forty-year review of extraction frequencies at a university orthodontic clinic. Angle Orthod. 1994;64:407–14.
2. Bramante MA. Controversies in orthodontics. Dent Clin N Am. 1990;34:91–102.
3. Ackerman JL, Proffit WR, Sarver DM. The emerging soft tissue paradigm in orthodontic diagnosis and treatment planning. Clin Orthod Res. 1999;2:49–52.
4. Jackson TH, Guez C, Lin F-C, Proffit WR, Ko C-C. Extraction frequencies at a university orthodontic clinic in the 21st century: Demographic and diagnostic factors affecting the likelihood of extraction. Am J Orthod Dentofacial Orthop. 2017;151:456–62.
5. Baumrind S, Korn EL, Boyd RL, Maxwell R. The decision to extract: part II. Analysis of clinicians' stated reasons for extraction. Am J Orthod Dentofac Orthop. 1996;109:393–402.
6. Konstantonis D, Anthopoulou C, Makou M. Extraction decision and identification of treatment predictors in Class I malocclusions. Prog Orthod. 2013;14:47.

7. Kim DW, Kim H, Nam W, Kim HJ, Cha IH. Machine learning to predict the occurrence of bisphosphonate-related osteonecrosis of the jaw associated with dental extraction: a preliminary report. Bone. 2018;116:207–14.
8. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 2018;77:106–11.
9. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.
10. McMaster C, Liew D, Keith C, Aminian P, Frauman A. A machine-learning algorithm to optimise automated adverse drug reaction detection from clinical coding. Drug Saf. 2019;42:721–5.
11. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38:1805–14.
12. Jung SK, Kim TW. New approach for the diagnosis of extractions with neural network machine learning. Am J Orthod Dentofac Orthop. 2016;149:127–33.
13. Li P, Kong D, Tang T, Su D, Yang P, Wang H, et al. Orthodontic treatment planning based on artificial neural networks. Sci Rep. 2019;9:2037.
14. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
15. Kavzoglu T. Chapter 33 – Object-oriented random forest for high resolution land cover mapping using quickbird-2 imagery. In: Samui P, Sekhar S, Balas VE, editors. Handbook of neural computation. London: Academic Press; 2017. p. 607–19.

8 Machine (Deep) Learning for Characterization of Craniofacial Variations

Si Chen, Te-Ju Wu, Tai-Hsien Wu, Matthew Pastewait, Anna Zheng, Li Wang, Xiaoyu Wang, and Ching-Chang Ko

8.1 Introduction

Craniofacial morphology variation is a common finding in the population. The severity of the variation can be used as a clue to reveal any potential underlying disease. Mild-to-moderate variations are frequently seen in orthodontic patients. Many malocclusions are associated with abnormal growth of the maxilla, mandible, or both. Over-/under-/asymmetrically developed maxillofacial bones result in a malformed skeletal base for the alignment of teeth and occlusion. For example, impacted teeth can be the result of deficient skeletal space. Severe variations are usually seen in craniofacial anomalies (CFA), a group of various deformities in the growth of the head and facial bones. Patients with CFA have a different appearance from normal. These congenital abnormalities are present at birth and have numerous variations of different severity. Cleft lip and/or palate, a separation/gap that happens in the upper lip and/or the palate, is the most common congenital CFA seen at birth, occurring in approximately 1/1000 births.

Many factors may contribute to the development of craniofacial variations, among which gene mutation and environment influence are the most frequent. A change in the genes at the time of conception can result in a craniofacial variation. Environmental exposure to some risk factors, such as tobacco consumption or folic acid deficiency during pregnancy, may play a role in the development of cleft lip and/or palate.

S. Chen (✉)
Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing, China

T.-J. Wu
Department of Orthodontics, Kaohsiung Chang Gung Memorial Hospital, Kaohsuing, Taiwan

T.-H. Wu · C.-C. Ko
Division of Orthodontics, College of Dentistry, The Ohio State University, Columbus, OH, USA

M. Pastewait
United States Air Force, Kadena AB, Okinawa, Japan

A. Zheng
Division of Oral and Craniofacial Health Sciences, Adam School of Dentisty, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

L. Wang
Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

X. Wang
Department of Orthodontics, Beijing Stomatological Hospital, Captial Medical University, Beijing, China

Department of Stomatology, Beijing Tian Tan Hospital, Capital Medical University, Beijing, China

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_8

Impacted maxillary canines are found in approximately 2% of the general population, and 83–92% of all impacted maxillary canine cases are unilateral impactions. While the exact etiology is unknown, it has been proposed that localized, systemic, or genetic etiologic factors contribute to canine impaction. Because maxillary canines have the longest duration of development in the deepest area of the maxilla, followed by the longest and most tortuous path of eruption, skeletal maxillary structure variation may be either an etiologic factor affecting canine eruption or a result of underdevelopment of the maxilla.

Cleft lip is an abnormality in which the lip does not completely form at the 7th week of embryonic development. It occurs when the medial nasal prominence and maxillary prominence fail to fuse. The degree of the cleft lip can vary greatly, from mild (notching of the lip) to severe (large opening from the lip up through the nose). The cleft palate happens when the roof of the mouth does not completely close, resulting from the failure of the palatal shelves to fuse with each other in the horizontal plane at the 12th week of embryonic development and leaving an opening that can extend into the nasal cavity. The cleft can involve either side or both sides of the palate. It can extend from the alveolar ridge of the hard palate to the soft palate. It can occur in isolation or in combination with cleft lip.

Early diagnosis of craniofacial variations is important because malformation of the hard and soft tissues can cause both functional and aesthetic problems from an early stage. Accurate quantitative analysis of the morphological deformities will be very useful in the treatment of these patients, especially for planning craniofacial surgery and orthodontic treatment. The plastic/craniofacial surgeon performs surgery to correct the abnormalities of the skull, facial bones, and soft tissue. The orthodontist evaluates tooth positions and corrects the malocclusion to improve the oral function.

Radiographic imaging is routinely used for the diagnostic evaluation and treatment planning of patients with craniofacial variations. In the last decade, cone beam computed tomography (CBCT) has received increased attention due to its high space accuracy and low radiation exposure and cost. CBCT is now widely used as an alternative to spiral computed tomography (CT) in locating anatomical structures and visualizing craniofacial abnormalities. To accurately assess craniofacial deformities, one critical step is to segment the CBCT image in order to generate a three-dimensional (3D) model, which includes segmentation of bony structures from soft tissues and separation of the mandible from the maxilla. However, due to the image artifacts caused by beam hardening, imaging noise, inhomogeneity, and truncation, it is very difficult to segment the CBCT based on thresholding and morphological operations automatically. In addition, manual segmentation is tedious, time-consuming, and user-dependent. It may take 5–6 h to segment the maxillomandibular region.

Machine learning is a branch of artificial intelligence that uncovers patterns in data automatically and then applies the detected patterns for future data prediction or decision making. This field has exploded in development as a result of rapid increases in computer storage capacity and processing power, thereby allowing for the application of machine learning in medical diagnosis, bioinformatics, robotics, and more. Preliminary applications of machine learning in dentistry mainly involve multiple variable data analyses using modern computationally intensive methods. Kim et al. applied machine learning to predict the occurrence of bisphosphonate-related osteonecrosis of the jaw associated with dental extraction [1]. Montenegro et al. applied machine learning techniques for caries prediction [2]. Hammond et al. used computational models to address problems in the development, growth, and repair of oral and craniofacial tissues arising in cell biology, clinical genetics, and dentistry [3]. Murata et al. proposed a deep learning model to provide an objective morphological assessment of facial features for orthodontic treatment [4].

For patients with craniofacial variations, machine learning can be a very useful tool to study the morphological differences between healthy and abnormal populations. A good example is the application of machine learning in the automatic segmentation of CBCT images. Regions of inter-

est in the craniofacial area for both normal and pathological subjects can be segmented in a far more efficient way using a learning-based framework. Therefore, 3D morphological deviations of these patients can be studied in a more precise and quantitative way.

In this chapter, we first introduce a method called Reverse Reconstruction (RR) using computer-aided design (CAD) technology for preparation of the ground truth for image segmentation. Taking cleft palate as an example, which is challenging to segment manually from scratch, RR facilitates the data preparation before training a machine (deep) learning model. Then, we comprehensively introduce how to implement the automatic segmentation of CBCT images using a well-known deep learning method, called 3D UNet, for automatic segmentation of maxillae and cleft defects. After that, a published application of machine learning in patients with craniofacial variation is reviewed in which an automatic 3D segmentation method was used to examine maxilla morphology variations in patients with impacted canines.

Machine learning holds great promise for treating patients with craniofacial variations. With regard to cleft palate patients, artificial intelligence-assisted evaluation of the alveolar cleft volume and its architecture will not only benefit the current autograft procedure but also provide a precedent toward tissue engineering for customized scaffold construction. In general, digital treatment planning and surgical simulation can be performed based on individual models of virtual reality. Machine learning will provide patients with craniofacial variations the opportunity for more predictable and promising treatment results.

8.2 Automatic Three-Dimensional Image Segmentation

One of the major advantages of machine learning is to help expedite time-consuming segmentation tasks. Currently, image recognition and differentiation by artificial intelligence are widely used for assisted diagnosis. The researchers label the regions of interest, which possess diagnostic information. All of these collected images with corresponding labels (known as the "ground truth") are used as the training dataset. This dataset helps to build algorithms that can learn from and make predictions on data. Many research groups have used machine learning with plain films to predict or rank pathological changes. However, building up the ground truth in three dimensions is another story.

Segmentation of CT or CBCT images has been widely used in dentistry, which is a crucial step for generating 3D models for advanced diagnosis, surgery, and treatment planning. Segmentation methods can be divided into three categories: *manual segmentation*, *semiautomatic segmentation*, and *automatic segmentation*. Routinely, the practitioner spends a lot of time tracing objects of interest and isolating objects from 3D images, slice by slice [5, 6]. This category is the so-called *manual segmentation*.

Since manual segmentation is time-consuming and tedious, image processing algorithms have been developed to determine the boundaries among different categories of targets automatically. *Semiautomatic segmentation* means that some algorithms (e.g., seeded region growing SRG [7]) require initial conditions or settings (e.g., seed in SRG) given by a user. *Automatic segmentation* is based on machine learning approaches with numerous training samples for learning how to determine the boundaries among different categories of targets. Once a well-trained model is obtained, it can predict segmentation results on the new unseen images without human intervention. However, machine learning requires abundant annotated data (i.e., images and ground truth), which is usually obtained using manual segmentation, semiautomatic segmentation, or a combination of both methods (e.g., a subtractive method based on software Mimics [8]). In this section, we illustrate how to prepare samples with ground truth efficiently and then how to implement learning-based automatic segmentation methods.

8.2.1 Preparing Training Samples and Ground Truth

To prepare the ground truth of training samples (e.g., CBCT images), we suggest two open-source graphical user interface (GUI) programs: 3D Slicer (www.slicer.org) and ITK-SNAP (www.itksnap.org) [9]. Both programs have many effective functions for manual and semiautomatic segmentations. To learn how to use 3D Slicer and ITK-SNAP in detail, please refer to their websites.

While semiautomatic segmentation methods help solve time-consuming issues of segmentation, there are other difficulties in segmenting maxillofacial structures in dentistry. First, the similar radiodensity, or grayscale, between the target object and its surrounding structures creates a problem of differentiation. For example, it can be difficult to isolate a dental root because of the radio-similarity (gray level value) of cementum and alveolar bone. Although using higher resolution radiographs would help to minimize this problem, the overexposure to radiation raises concerns. In addition, it takes more time to process the increased number of slices.

Second, for those structures without identifiable borders, practitioners have difficulties outlining structure contours through all of the slices. For instance, it is challenging to delineate the defect outline in cleft palate patients. The defects of a cleft palate typically have four boundaries facing the soft tissue outline, including front (alveolar), rear, nasal, and palatal borders. For the palatal border, practitioners can use the neighboring bony structure to predict the contour of the cleft. However, the remaining borders are difficult to outline.

To address such difficulties, reverse processing using CAD can provide a better method to distinguish the geometrical shapes of various defects. Here, we use the cleft palate as an example, and present a method called Reverse Reconstruction (RR), which uses CAD software (e.g., Rhinoceros 5.0 (Robert McNeel & Associates, Seattle, Wash.)) to make the ground truth. In other words, the defects of interest are outlined first as 3D rendering objects, followed by proper cropping to present the actual defects. The detailed procedures of the RR method are described as follows:

Step 1: Based on a Digital Imaging and Communications in Medicine (DICOM) file, use a threshold-based method to segment the maxilla, and then export to a stereolithography (STL) file (let us name it *maxilla.stl*). This process can be done by using both 3D Slicer and ITK-SNAP. Next, import the *maxilla.stl* in the CAD software. Steps 2 to 7 are carried out in the CAD software.

Step 2: Use a 3D spline to outline the nasal border of the defect by "bridging" the neighboring bony points along the nasal lining, as shown in Fig. 8.1.

Step 3: Use a 3D spline to outline the palatal border of the defect along the palatal vault, as shown in Fig. 8.2.

Step 4: Outline the frontal border of the defect along the curvature of the frontal alveolar bone, as shown in Fig. 8.3.

Step 5: Smooth the connected surfaces to delineate the proper outline in the shape, as shown in Fig. 8.4.

Step 6: The margin in the rear of the defect is evaluated according to the symmetry of the whole maxilla. The mirror image of the healthy site

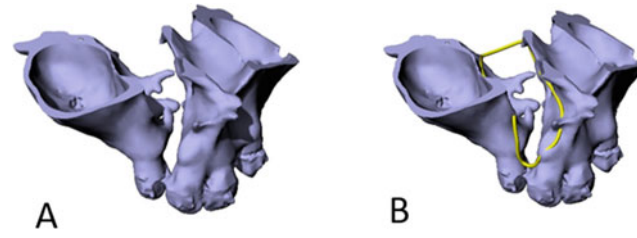


Fig. 8.1 (**a**) The practitioners outline the ceiling of the defect. (**b**) The complete outlining of the ceiling of the defect

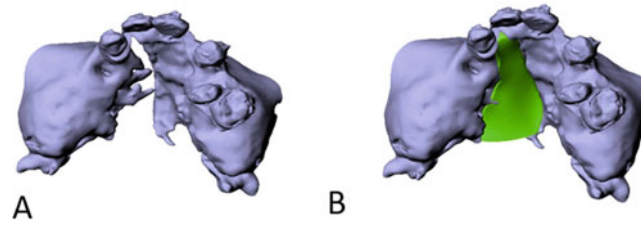


Fig. 8.2 (**a**) The practitioners outline the base of the defect. (**b**) The complete outlining of the base of the defect

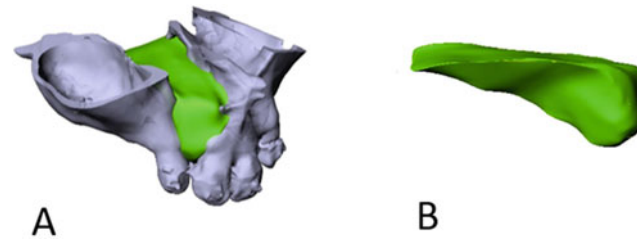


Fig. 8.3 (**a**) The frontal contour of the defect is depicted according to the symmetry and smoothness of the alveolar surface. (**b**) The initial defect 3D model

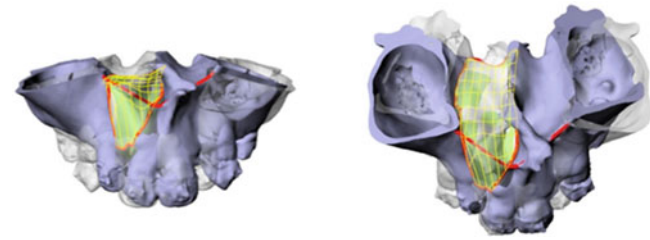


Fig. 8.4 The nasal contour was outlined with proper curvature (the yellow line)

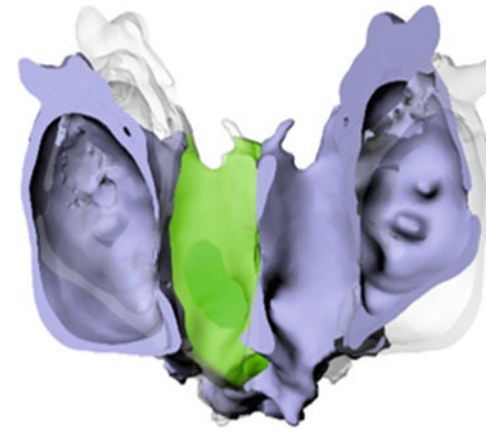


Fig. 8.5 The mirror image of uncleft site (the gray object) were used as the reference to define the posterior border of cleft

could be used as a reference, as shown in Fig. 8.5.

Step 7: Use the Boolean operation to crop out the isolated defect in the shape, as shown in Fig. 8.6.

After filling in the cleft frame, the whole object (i.e., the cleft defect) can be exported as an STL file (let us name it *cleft.stl*). We will illustrate how to voxelize the 3D model of the cleft defect (i.e., *cleft.stl*) using 3D slicer, as follows:

Step 8: Load *maxilla.stl* and *cleft.stl* in 3D slicer as *Model* mode, as shown in Fig. 8.7.

Step 9: Right-click on the *maxilla* model and select *Convert model to segmentation node*, as shown in Fig. 8.8a. Repeat the same process on *cleft* model. We can see two *segmentation nodes* appeared under *Data module.*

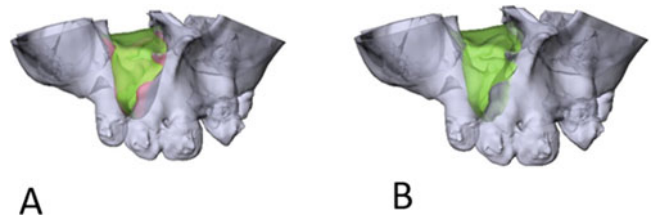


Fig. 8.6 (**a**) The collision parts (colored in red) between the maxilla and the reconstructed defect two 3D objects were removed after Boolean operation. (**b**) The final 3D models of maxilla and defect

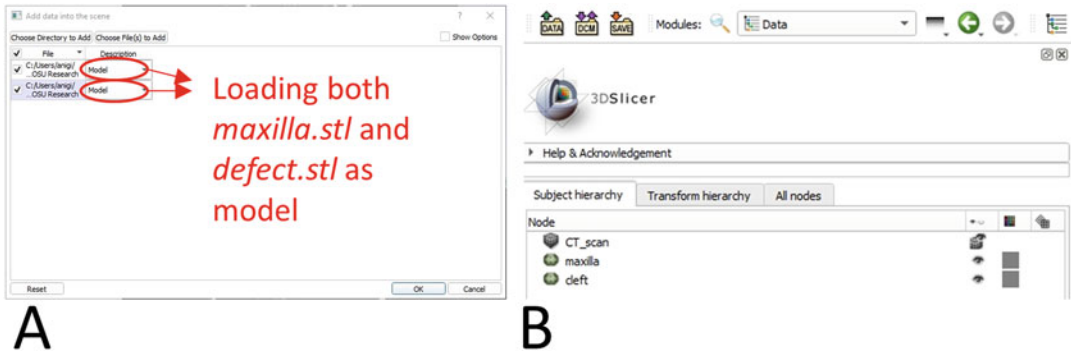


Fig. 8.7 (**a**) When loading *maxilla.stl* and *defect.stl* in 3D slicer, select *Model* in *Description*. (**b**) The two models (*maxilla* and *cleft* in the figure) will appear under *Data module* in 3D Slicer

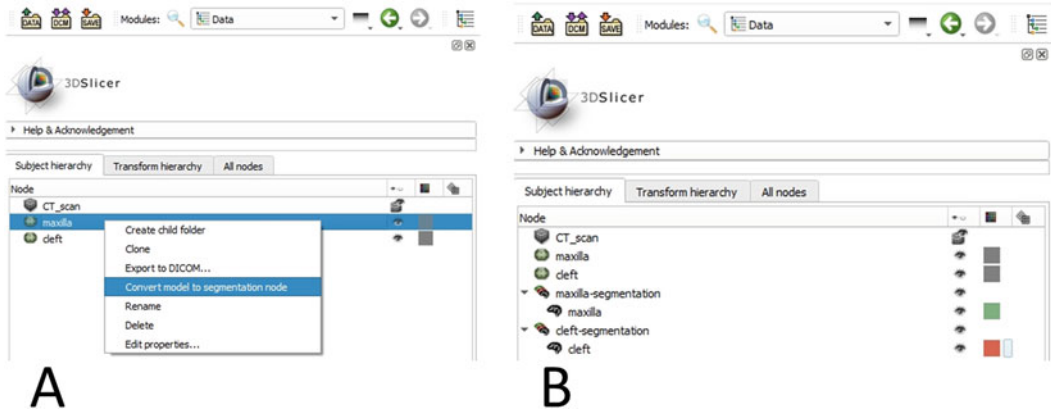


Fig. 8.8 (**a**) Converting models to segmentation nodes. (**b**) After that, we can see two *segmentation* nodes appeared under *Data moudule*. We can change the color of *cleft-segmentation* node to distinguish the *maxilla-segmentation* and *cleft-segmentation* nodes

We can change the color of *cleft-segmentation* node (e.g., red) to distinguish the *maxilla-segmentation* and *cleft-segmentation* nodes, as shown in Fig. 8.8b.

Step 10: Now, merge the two segmentation nodes. Right-click on the *cleft-segmentation* node and select *Edit properties*, as shown in Fig. 8.9a. In *Copy/move segments* section (see highlighted red circle), first set *Current segmentation* as m*axilla-segmentation* (highlighted red circle), then select *cleft*, and finally click ">" (highlighted red circle). After that, we can

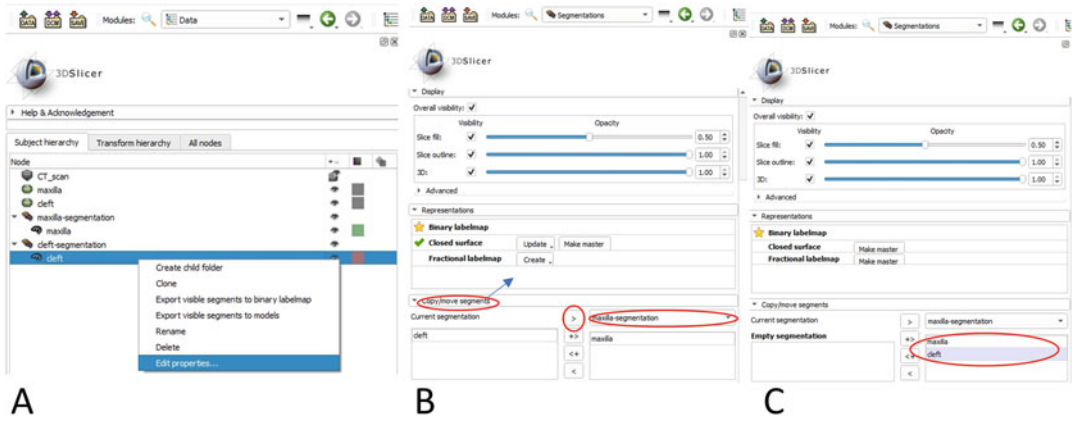


Fig. 8.9 (**a**) Right-click on the *cleft-segmentation* node and select *Edit properties*. (**b**) In *Copy/move segments* section (see highlighted red circle), first set *Current segmentation* as m*axilla-segmentation* (highlighted red circle), then select *cleft*, and finally click "*>*". (**c**) The *cleft* will move under *maxilla-segmentation*

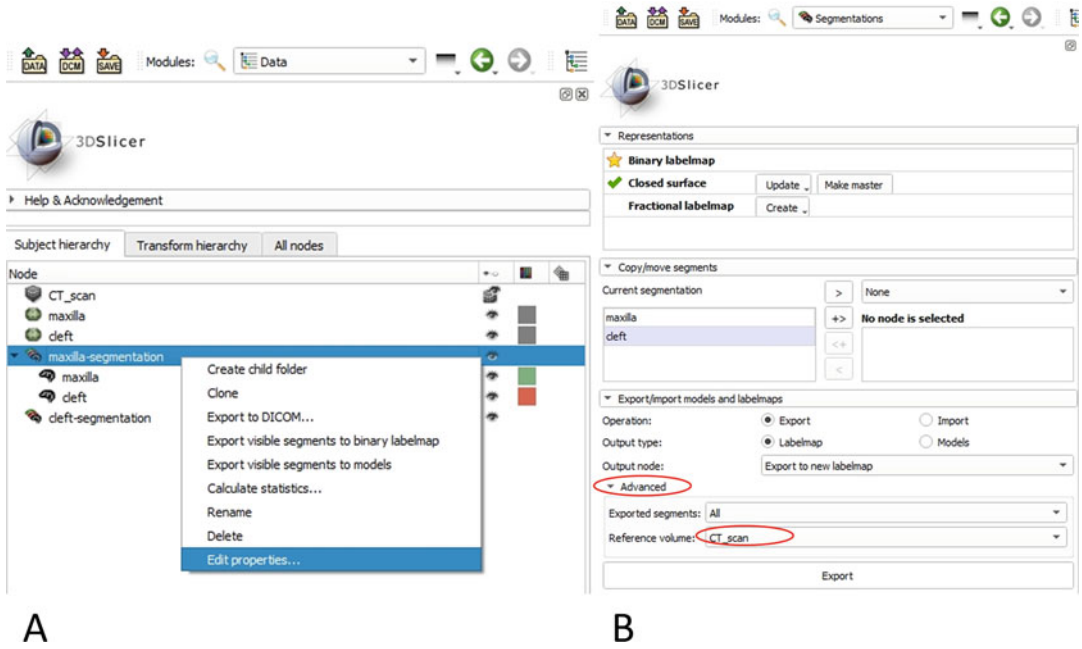


Fig. 8.10 (**a**) Right-click on the *maxilla-segmentation* and select *Edit properties*. (**b**) In the *Export/import models and labelmaps* section, expand *Advanced* (see highlighted red circle), set *Reference volume* to the original CT scan (see highlighted red circle), and then click *Export*

see both *maxilla* and *cleft* are under *maxilla-segmentation* node (see Fig. 8.10a).

Step 11: Convert the segmentation node (i.e., *maxilla-segmentation*) into *labelmap*. This process will voxelize the 3D models to volumetric data like DICOM, serving as the ground truth of the DICOM images (i.e., CT/CBCT scans) for machine learning. To do that, right-click on the *maxilla-segmentation* and select *Edit properties*. In the *Export/import models and labelmaps* section, expand *Advanced* (see highlighted red

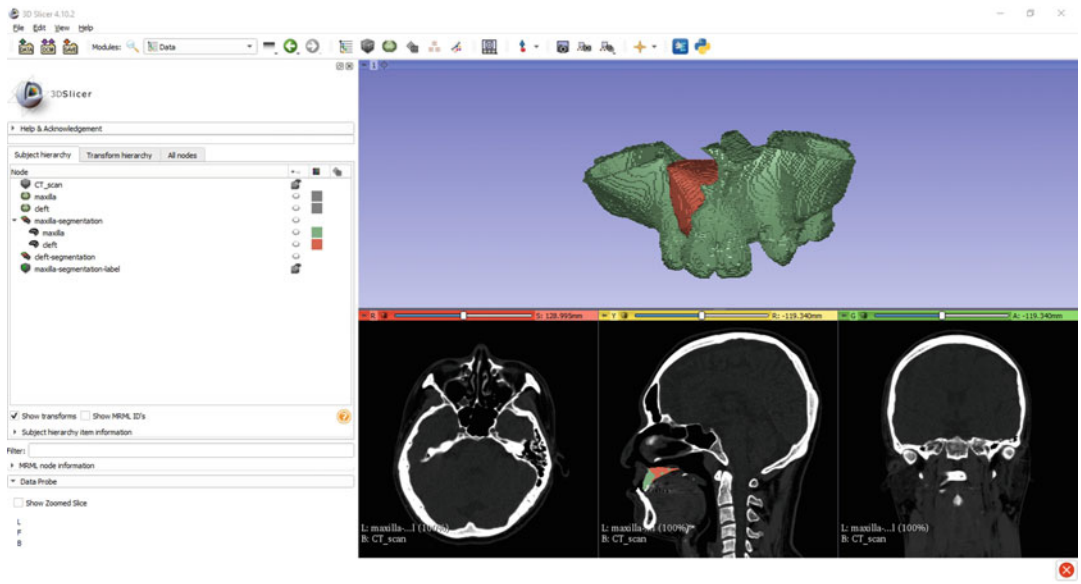


Fig. 8.11 Original DICOM file (i.e., CT/CBCT scan) with the segmentation generated from the RR method

circle), set *Reference volume* to the original CT scan (see highlighted red circle), and then click *Export*. Figure 8.11 shows the final outcome.

8.2.2 A Machine Learning-Based Method: 3D U-Net

At this point, we assume that we have well-prepared, annotated training data that is ready for training a machine/deep learning model. Next, we introduce U-Net, which is a well-known deep learning neural network architecture for biomedical image segmentation, developed by Ronneberger et al. in 2015 [10]. Later, Çiçek et al. further expanded this network architecture for 3D volumetric segmentation [11].

Recently, deep convolutional networks have achieved dramatic success in image recognition. The two main reasons involve deeper neural networks and increasing amounts of available data. However, obtaining a large number of training images is very difficult in the biomedical field. Borrowing an idea from [12], U-Net uses patches (i.e., small pieces of an image) as the training data. In general, a patch is rectangular, and an image can generate numerous patches. Thus, the amount of training data in terms of patches is much larger than the number of original images.

Although the use of patches can overcome the shortage of training data in biomedicine, using patches has a tradeoff problem between context information and localization. A large size patch requires more max-pooling layers that reduce the localization accuracy, whereas a small size patch results in less context information. Below, we introduce the 3D U-Net architecture and explain how to use 3D U-Net to address this tradeoff problem between context information and localization.

Figure 8.12 illustrates the architecture of 3D U-Net, which has two parts, a contracting path and an expansive path. The contracting path consists of four layers. Each layer is composed of two $3 \times 3 \times 3$ convolutional layers each followed by a rectified linear unit (ReLU) activation function, and then a $2 \times 2 \times 2$ max pooling layer with strides of two for downsampling. The number of channels of each convolutional layer is doubled before max pooling. The expanding path also consists of 4 layers. Each of these layersis com-

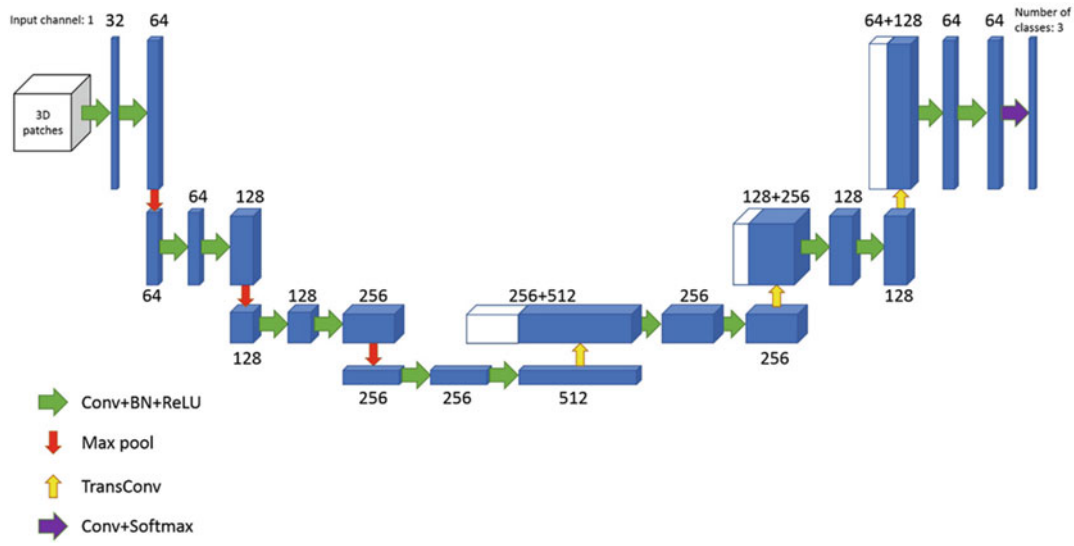


Fig. 8.12 3D U-Net architecture. Reprinted from [11] with permission the Springer Nature

posed of an up-convolutional layer with stride 2, followed by a concatenation operation with the corresponding feature map from the contracting path, and two $3 \times 3 \times 3$ convolutional layers, each followed by a ReLU function. In the last layer, a $1 \times 1 \times 1$ convolutional layer reduces the number of output channels to the number of labels, followed by a softmax layer. Note that a batch normalization is used before each ReLU in the entire network.

The first contracting path (left part) is used to capture the context in the image. For those who are familiar with the convolutional neural network (CNNs), the contracting path is an encoder and similar to the classic CNN structure in classification problems. The second expansive path (right part) is a decoder and allows precise localization using transposed convolutions. The outputs of feature maps from each layer in the contracting path are concatenated to the corresponding layer in the expansive path (the white boxes shown in Fig. 8.12), which is called skip connection. These skip connections provide essential features learned from the contracting path into the expansive path to recover the full spatial information. Based on its special symmetric architecture, 3D U-Net demonstrates its outstanding performance in image segmentation.

8.2.3 Segmentation Results of Maxillae and Palatal Defect Based on 3D UNet

Using 3D UNet, we conducted a study to characterize 3D morphometry of the maxilla and the cleft defect in non-syndromic patients with unilateral cleft lip and palate (UCLP). A model was trained and tested using 30 CBCT images and their manual segmentations, where 24 images served as the training set, 3 images served as a validation set, and the rest three were the test set. All CBCT data were acquired on the same CBCT scanner (NewTom, Verona, Italy) (110 kV, 1–20 mA, 15×15 cm field of view, 0.250 mm voxel size) from the Department of Orthodontics, Beijing Stomatological Hospital, Capital Medical University, under the institutional review board (IRB: KY2017-072-01). This segmentation task was a voxel-wise classification problem with three classes corresponding to maxillae, cleft defects, and background. In the training set, each CBCT image generated 1500 patches with a size of $64 \times 64 \times 64$, so the total number of the training dataset was 36,000 patches. In the test set, we fed the original CBCT images to the model and compared the predicted automatic segmentation with the manual segmentation.

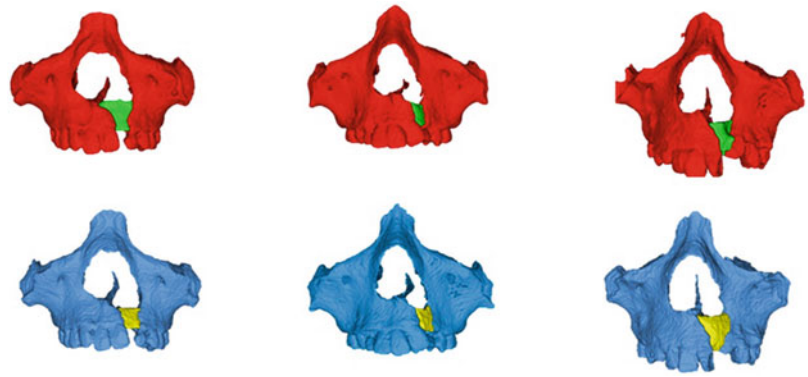


Fig. 8.13 The segmentation results in the test samples. The first and second rows represent the results from 3D UNet and manual, respectively. The red and blue parts represent the maxillae, and the green and yellow parts represent the cleft defects

The model was optimized to have minimal generalized Dice loss [13] for 50 epochs using the Adam optimizer [14]. All processes were implemented using PyTorch, an open-source machine learning library in Python computer language (pytorch.org).

The Dice similarity coefficient (DSC) is a standard index to evaluate the similarity between two images. The definition of the DSC is expressed as

$$\mathrm{DSC} = \frac{2\,|A \cap B|}{|A| + |B|},$$

where $|A|$ and $|B|$ represent the cardinalities of the learned and manual sets, and $|A \cap B|$ represents the intersection of the two sets. A value of 0 indicates no similarity, whereas a value of 1 indicates perfect agreement. The average DSCs of the maxilla and defect between the manual and automatic segmentation in the test samples were 0.92 ± 0.02 and 0.76 ± 0.05, respectively. This data suggests that the segmentations were accurate. Figure 8.13 shows that the auto-segmented maxilla and defect (red and green parts in the first row) were observed to have features similar to those obtained from manual segmentation (blue and yellow parts in the second row).

8.3 Application Review

In this section, we review a recent CFA application using machine learning: *unilateral impacted canine-related maxilla morphology variation,* [15]. Chen et al. used the random forest algorithm as the automated segmentation method. Although this method is different than the one we introduced above (i.e., 3D U-Net), the role of machine/deep learning to perform automated segmentation has no fundamental difference. Thus, the clinical findings are not significantly affected by the segmentation method. Below, we discuss the clinical findings of the application.

8.3.1 Unilateral Impacted Canine-Related Maxilla Morphology Variation

Chen et al. used machine learning to assess maxillary constriction in unilaterally impacted canine patients [15]. In their study, 60 patients were divided into two groups: 30 study group patients with unilaterally impacted canines and 30 healthy control group patients. The demographic distribution of study and control groups are given in Table 8.1. First, they used a series of random forest classifiers to segment maxillae. Then, they used an automated random forest-based landmark finder to propose three landmarks: basion (Ba), nasion (Na), and anterior nasal spine (ANS). These three landmarks defined a plane that was used to divide the maxillary segmentation into two sides. In their study, the average Dice similarity coefficient was 0.800 ± 0.027, ranging from 0.742 to 0.830, for automated segmentation, as shown in Fig. 8.14a. In addition, the mean differences in voxel position between the manually and automatically identified landmarks were 1.92 ± 1.02 for Ba,

Table 8.1 The demographic distributions of Study Group and Control Group

	Study Group	Control Group
Number of application samples	30	30
Female, *n* (%)	12 (40)	18 (60)
Mean age ± SD, year	14.97 ±2.04	14.53±2.24
Age range, year	11–18	11–18
Number of buccal, *n* (%)	16 (53)	N/A
Number of mid-alveolus, *n* (%)	4 (13)	N/A
Number of palatal, *n* (%)	10 (33)	N/A

SD indicates standard deviation; N/A indicates not applicable
Reprinted from [15] with permission from the E. H. Angle Education and Research Foundation

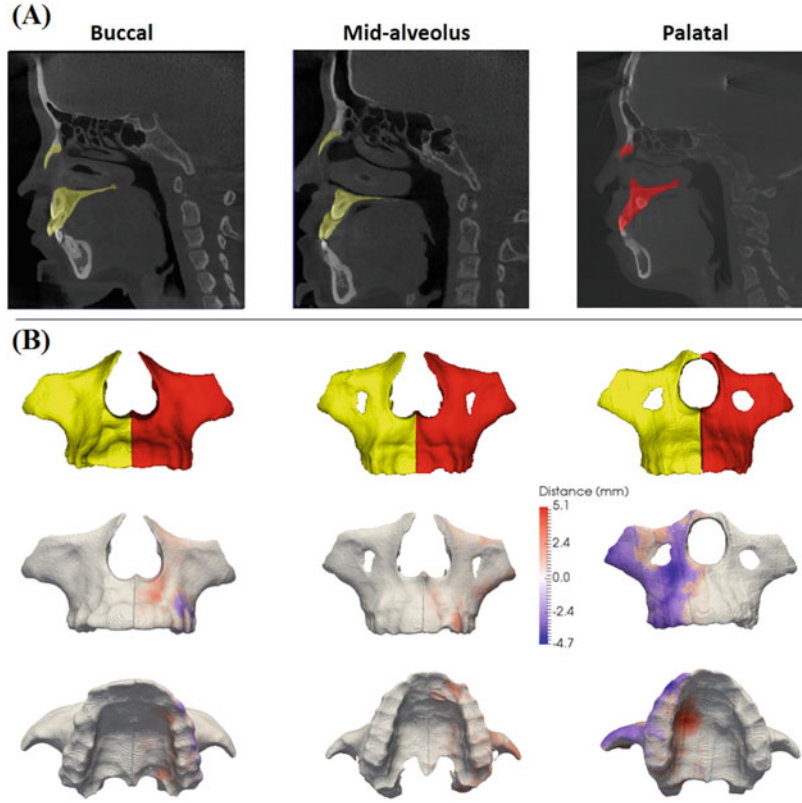


Fig. 8.14 (**a**) Segmentation results for the maxilla. (**b**) The superimposition results of three different types of impaction (buccal, mid-alveolus, and palatal) are shown, allowing for geometric difference determination. Reprinted from [15] with permission the E. H. Angle Education and Research Foundation

2.23 ± 1.19 for Na, and 2.26 ± 1.38 for ANS, where 1 voxel equals 0.3 mm.

The clinical outcomes showed that, on average, the maxillary volume of the study group tended to be smaller (5000 mm^3 less) than the control group (p-value = 0.006). More specifically, the comparison between the maxillary volume of males to females, demonstrated that the maxillary bone volume in males ($5.36 \pm 0.71 \times 10^4$ mm^3) was significantly greater than in females ($4.59 \pm 0.44 \times 10^4$ mm^3) (p-value <0.001). Even after the adjustment for age and gender, the differences in maxillary volume between the study group ($4.73 \pm 0.67 \times 10^4$

mm^3) and control group ($5.22 \pm 0.65 \times 10^4$ mm^3) were still significant (p-value = 0.023). However, when specifically comparing the volumetric differences between the impacted side ($2.37 \pm 0.34 \times 10^4$ mm^3) and the non-impacted side ($2.36 \pm 0.35 \times 10^4$ mm^3) of the maxilla in the study group alone, the differences were not significant.

In addition, three linear measurements (maxillary width, height, and depth) were defined and compared between the two groups. The definitions of the maxillary width, height, and depth were distances measured between left and right jugular points, distances measured between the most superior border of the frontal process of the maxilla and the most inferior point of the alveolar process, and distances measured from ANS to a point directly posterior to ANS, respectively. The average width, height, and depth in males (67.4 ± 4.4, 67.1 ± 3.4, and 50.8 ± 2.7 mm, respectively) were significantly greater than in females (63.5 ± 4.1, 64.9 ± 3.6, and 46.5 ± 3.0 mm, respectively) (p-value = 0.002, p-value = 0.019, and p-value <0.001). Furthermore, the study group also tended to have smaller width, height, and depth than the control group, as shown in Table 8.2. After adjusting for both age and sex, the impaction effect in all dimensions became non-significant.

The models created by automatic segmentation and landmark finder allow for analysis of the two sides of the maxilla. The superimposition maps illustrated in Fig. 8.14b represent the morphological discrepancies between the impaction and non-impaction maxillary halves. The maps show the superimposition of three different types of impacted canines: buccal, mid-alveolus, and palatal. The models demonstrate that palatally impacted canines appear to contain more discrepancies between their impaction and non-impaction sides than their counterparts, buccal and alveolar deformations. The maps detail that the two halves have approximately 2.4 mm of transverse constriction, which aligns better with the average difference of 2.3 mm from all 30 study group subjects than that of the control group subjects. In clinic, the 2 mm difference can be fixed with simple archwire expansion. Regardless, the volumetric constriction was found to be 5000 mm^3 which is approximately five times the volume of a canine. This data explains how maxillary underdevelopment is more likely when unilateral canine impaction is present. The data might support O'Neill's finding [16] where the use of rapid maxillary expansion in early mixed dentition effectively increased the rate of eruption of palatally displaced maxillary canines compared to an untreated control group.

Table 8.2 The clinical outcomes of Study Group and Control Group

	Study Group	Control Group
Bone volume Impacted side for study group Left side for control group Mean ± SD (10^4 mm^3)	2.36 ± 0.35 Max: 33.5; Min: 18.3	2.57 ± 0.30 Max: 36.6; Min: 22.4
Bone volume Non-impacted side for study group Right side for control group Mean ± SD* (10^4 mm^3)	2.37 ± 0.34 Max: 34.2; Min: 18.6	2.65 ± 0.38 Max: 31.4; Min: 21.2
Maxillary width (mm)	64.3 ± 5.3 Max: 78.8; Min: 56.1	66.6 ± 3.6 Max: 73.2; Min: 59.7
Maxillary height (mm)	65.1 ± 3.6 Max: 70.0; Min: 55.0	67.0 ± 3.5 Max: 74.6; Min: 59.7
Maxillary depth (mm)	47.7 ± 3.6 Max: 55.5; Min: 41.2	49.6 ± 3.3 Max: 56.4; Min: 40.7

Reprinted from [15] with permission the E. H. Angle Education and Research Foundation

8.3.2 Discussions

With the aid of CT/CBCT, both 3D volumetric and 2D linear measurements are achievable in dentistry, allowing for the assessment of craniofacial growth and more specialized dental diagnosis and treatment. However, manual segmentation becomes the bottleneck in clinical research because it is time-consuming and tedious. Although semiautomatic segmentation has improved the efficiency of segmentation, not all of the semiautomatic segmentation methods can consider the morphology of targets with intensities similar to surrounding structures (e.g., maxillae).

As artificial intelligence, or machine/deep learning, has achieved ever-increasing success in biomedical image processing, the learning-based method is a very suitable candidate for the segmentation of structures with complex morphology. In both CFA applications discussed above, it is obvious that the generated data would have been impossible without the implementation of the machine learning approach. In particular, the segmented targets of maxillae and cleft defects cannot be analyzed efficiently using semiautomatic segmentation methods. Nevertheless, the accuracy of automatically segmented targets needs to be improved. As machine learning continues to progress and develop, it is expected that the accuracy of segmentation will be satisfied even in the use of clinical practice, and the results will provide more informative evidence for clinical treatment.

8.4 Conclusion

Artificial intelligence, including machine learning and deep learning, is rapidly expanding into multiple facets of society. The field of dentistry can benefit by adapting to artificial intelligence for multiple reasons. First, patient encounters during treatment generate many types of data. Radiographs, digital photographs, intraoral and extraoral features are just a few types of data generated in the dental clinic. Artificial intelligence can perform analytics to decipher this information and aid in efficient diagnosis and treatment planning. This chapter presented one demonstration and one published application review where machine learning assisted in the segmentation and analyses of maxillae for two craniofacial anomalies, cleft palates, and impacted canines. The manual segmentation of thousands of images would have taken the typical researcher weeks, if not months, to complete. Machine learning greatly streamlined this process in terms of both time and cost. Second, standardization of practice in the field of dentistry is low compared to other areas of healthcare. A range of valid treatment options exists for any given case. Using artificial intelligence and large datasets (that include diagnostic results, treatments, and outcomes), one can now empirically measure the effectiveness of different treatment modalities given very specific clinical findings and conditions. Third, dentistry is largely practiced by independent clinicians in their own offices. These dentists have the autonomy to adopt beneficial technologies without the bureaucracy often found in large healthcare organizations. In order to remain competitive in the modern dental market, dentists must be proactive in seeking innovation and adopting various technologies. The U-Net neural network described above has been used in the segmentation of brain and liver images. Nevertheless, it has a practical use in dentistry, which also relies heavily on biomedical imaging. Despite the promise of artificial intelligence, the volume of dental research in this field is relatively low. Further, the clinical accuracy of artificial intelligence must be improved with an increased number and variety of cases. Before artificial intelligence can take on a more important role in making diagnostic recommendations, the volume and quality of research data will need to increase.

References

1. Kim DW, Kim H, Nam W, Kim HJ, Cha I-H. Machine learning to predict the occurrence of bisphosphonate-related osteonecrosis of the jaw associated with dental extraction: a preliminary report. Bone. 2018;116:207–14. https://doi.org/10.1016/j.bone.2018.04.020.
2. Montenegro RD, Oliveira ALI, Cabral GG, Katz CRT, Rosenblatt A. A comparative study of machine learning techniques for caries prediction. In: 2008 20th IEEE international conference on tools with artifi-

cial intelligence, 2008, vol. 2, pp. 477–481, https://doi.org/10.1109/ICTAI.2008.138.
3. Hammond P, Hutton T, Maheswaran S, Modgil S. Computational models of oral and craniofacial development, growth, and repair. Adv Dent Res. 2003;17(1):61–4. https://doi.org/10.1177/154407370301700114.
4. Murata S, Lee C, Tanikawa C, Date S. Towards a fully automated diagnostic system for orthodontic treatment in dentistry. In: 2017 IEEE 13th international conference on e-science (e-science), 2017, pp. 1–8, https://doi.org/10.1109/eScience.2017.12.
5. Gan Y, Xia Z, Xiong J, Zhao Q, Hu Y, Zhang J. Toward accurate tooth segmentation from computed tomography images using a hybrid level set model. Med Phys. 2015;42(1):14–27. https://doi.org/10.1118/1.4901521.
6. Wang L, et al. Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization. Med Phys. 2014;41(4):043503. https://doi.org/10.1118/1.4868455.
7. Adams R, Bischof L. Seeded region growing. IEEE Trans Pattern Anal Mach Intell. 1994;16(6):641–7. https://doi.org/10.1109/34.295913.
8. Chen GC, Sun M, Yin NB, Li HD. A novel method to calculate the volume of alveolar cleft defect before surgery. J Craniofac Surg., vol. 29, no. 2, 2018 [Online]. Available from https://journals.lww.com/jcraniofacialsurgery/Fulltext/2018/03000/A_Novel_Method_to_Calculate_the_Volume_of_Alveolar.20.aspx
9. Yushkevich PA, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage. 2006;31(3):1116–28. https://doi.org/10.1016/j.neuroimage.2006.01.015.
10. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation BT – Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015, pp. 234–241.
11. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation BT – Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, 2016, pp. 424–432.
12. Ciresan D, Giusti A, Gambardella LM, Schmidhuber J. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems, 2012, pp. 2843–2851.
13. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations BT – Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2017, pp. 240–248.
14. Reddi SJ, Kale S, Kumar S. On the convergence of Adam and beyond, eprint arXiv:1904.09237. p. arXiv:1904.09237, 01-Apr-2019 [Online]. Available from https://ui.adsabs.harvard.edu/abs/2019arXiv190409237R
15. Chen S, et al. Machine learning in orthodontics: Introducing a 3D auto-segmentation and auto-landmark finder of CBCT images to assess maxillary constriction in unilateral impacted canine patients. Angle Orthod. 2019;90(1):77–84. https://doi.org/10.2319/012919-59.1.
16. O'Neill J. Maxillary expansion as an interceptive treatment for impacted canines. Evid Based Dent. vol. 11, p. 86, Sep. 2010 [Online]. https://doi.org/10.1038/sj.ebd.6400742.

Patient-Specific Reference Model for Planning Orthognathic Surgery

9

Hannah H. Deng, Li Wang, Yi Ren, Jaime Gateno, Zhen Tang, Ken-Chung Chen, Chunfeng Lian, Steve Guofang Shen, Philip Kin Man Lee, Pew-Thian Yap, Dinggang Shen, and James J. Xia

9.1 Introduction

A large number of people require orthognathic surgery to correct their jaw deformities [1, 2], either congenital or acquired. The goal of orthognathic surgery is to restore deformed jaws[1] back to the normal shape.

The success of orthognathic surgery depends on both the surgeon's skill and a detailed surgical plan [3–6]. However, because of the complexity of craniomaxillofacial (CMF) anatomy and deformity, surgical planning is extremely difficult. The computer-aided surgical simulation (CASS) has become the standard of care for planning orthognathic surgery. In routine orthognathic surgical planning, a surgeon simulates the surgery by virtually cutting a three-dimensional (3D) model of a patient's skull into multiple bony segments, and moving them individually to a desired position mainly based cephalometric analysis and clinical measurement. This is done by first restoring the patient's right and left face back to symmetry, then adjusting each bony segment to a desired

[1] Please Note: Unless otherwise specified, in the following text, the term "jaws" represents both upper (or maxillary) and lower (or mandibular) jaws.

H. H. Deng · J. Gateno · Z. Tang · K.-C. Chen
Department of Oral and Maxillofacial Surgery, Houston Methodist Research Institute, Houston, TX, USA

L. Wang · Y. Ren · P.-T. Yap
Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

J. Gateno · J. J. Xia (✉)
Houston Methodist Academic Institute, and Department of Surgery (Oral and Maxillofacial Surgery), Weill Medical College of Cornell University, Houston, TX, USA
e-mail: JXia@houstonmethodist.org

C. Lian
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, China

S. G. Shen
Department of Oral and Craniomaxillofacial Surgery, Shanghai Jiaotong University College of Medicine, Shanghai, China

P. K. M. Lee
Hong Kong Dental Implant and Maxillofacial Center, Central, Hong Kong

D. Shen
Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_9

position by comparing the patient's cephalometric measurements to the norms, which are averaged values taken from a group of gender- and ethnicity-based "normal" population. The problem with this method is that each normal value has a mean and a range and thus does not necessarily represent a specific patient's CMF geometry. Each face is different, thus the surgeon's experience and artistic talent can greatly affect the outcome of orthognathic surgery.

We have developed a novel method to automatically estimate the patient-specific normal jaw shape. We believe that if a surgeon can foresee what the normal anatomy of the patient should be, i.e., with a patient-specific and anatomically correct reference model, the surgical planning will be objective and personalized. The estimated jaw shape is used as the reference model for a surgeon tailoring a patient-specific surgical plan for orthognathic surgery. In this chapter, we present a method to automatically estimate patient-specific and anatomically correct reference models for planning orthognathic surgery. This method corrects the patients with non-syndromic jaw deformities, i.e., the midface does not require surgical correction and only the maxilla and mandible are involved in the surgery. An example of a patient with jaw deformity is shown in Fig. 9.1. The midface is anatomically correct, while upper and lower jaws (i.e., the maxilla and mandible) require a double-jaw orthognathic surgery.

The correct morphology of the midface is used as a prior to predict the normal jaw shape for patients with jaw deformities. We predict the normal jaw shape using a data-driven method based on Sparse Shape Composition (SSC) [7, 8]. First, we employ the sparse representation technique [9, 10] to represent the patients' midface models using those of normal subjects. Then we reconstruct the patient-specific jaw models using the derived sparse coefficients and jaw models from normal subjects. Last, we evaluate the outcome of the proposed method on both synthetic and real patient data. The results showed that our proposed method is able to reconstruct the normal patient-specific jaw models which are within the normal range quantitatively.

The computer-aided surgical planning method has been significantly improved in the past decades. Vannier et al. [11] proposed a computer-aided method to delineate abnormal facial soft tissue and bony morphology from computed tomography (CT) scans for craniofacial surgical procedure-planning and evaluation. Xia et al. [12] developed CASS system and clinical protocols for CMF surgery planning. The 3D models of both skull and face were generated and quantitatively analyzed for diagnosis. The surgery process was performed on the 3D models to predict the outcome for the patients. Zachow et al. [13] proposed a statistical 3D shape model of the human mandible for surgical reconstruction

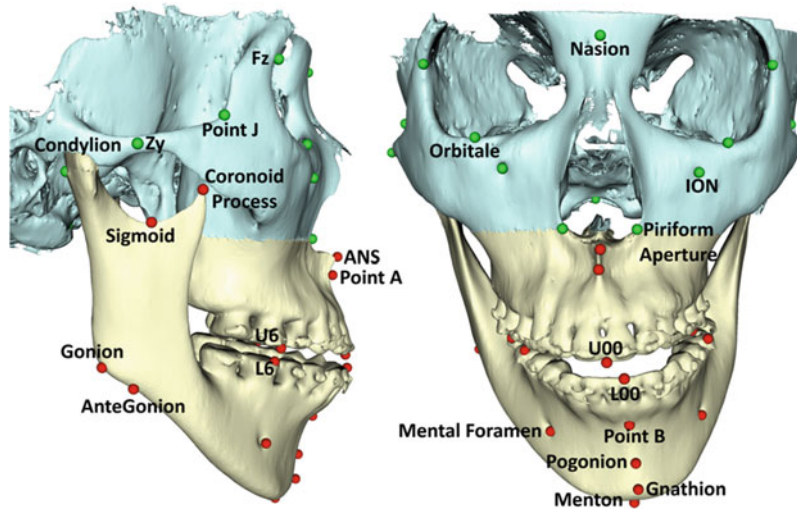


Fig. 9.1 Anatomical landmarks and surface models of a patient requiring orthognathic surgery. The skull surface is divided into two parts: the midface and the jaws. Midface model S_{mid} is marked in cyan and the abnormal jaw surface S_{jaw} is marked in yellow. The landmarks are partitioned into two groups: midface landmarks L_{mid} (green spheres) and jaw landmarks L_{jaw} (red spheres)

of bone defects. However, to the best of our knowledge, there is no existing surgical planning system that enables the prediction of and patient-specific [14, 15] anatomically correct reference model for CMF surgery.

9.2 Method

The flowchart of our proposed method is shown in Fig. 9.2. After acquiring the CT scan of a jaw deformity patient, we segment the CT image and generate the midface and jaw surface models, represented by S_{mid} and S_{jaw}, respectively. The midface model S_{mid} is considered normal and thus is not changed during surgery. Only the jaw model S_{jaw} needs to be corrected. Based on S_{mid}, we estimate a patient-specific jaw reference model to represent the normal jaw shape, which can be used as a guide to perform surgical planning on S_{jaw}. First, based on the normal midface database, we use sparse representation to approximately align the normal subjects onto the patient space. Then we use sparse coefficients and normal jaw models in the database to reconstruct the patient-specific normal jaw model. Finally, we combine the estimated normal jaw model and the patient's midface model together as the patient-specific CMF reference model.

The midface and jaw surface models are inefficient in calculation. Thus, we digitize anatomical landmarks on these models to effectively represent the models and improve the computation efficiency. These landmarks are digitized on both midface and jaw surface models positioned in the reference system [16], shown in Fig. 9.1. These landmarks are a subset selected from the anatomy landmarks to be both effective and efficient in jaw shape estimation. Specifically, it is computationally efficient to calculate based on a small subset of landmarks instead of the whole model, which includes tens of thousands of vertices and the mesh constructed by these vertices. In addition, all of the landmarks have clear clinical definitions, which are correlated to anatomic structure and shared among all patients. Last, it should be noted that all of the clinical cephalometric analyses are based on landmarks. Therefore, using landmarks in the jaw shape estimation is clinically relevant.

As shown in Fig. 9.1, there are 58 landmarks digitized on the skull model, among which 31 landmarks are digitized on the jaw and 27 are digitized on the midface. The clinical name is

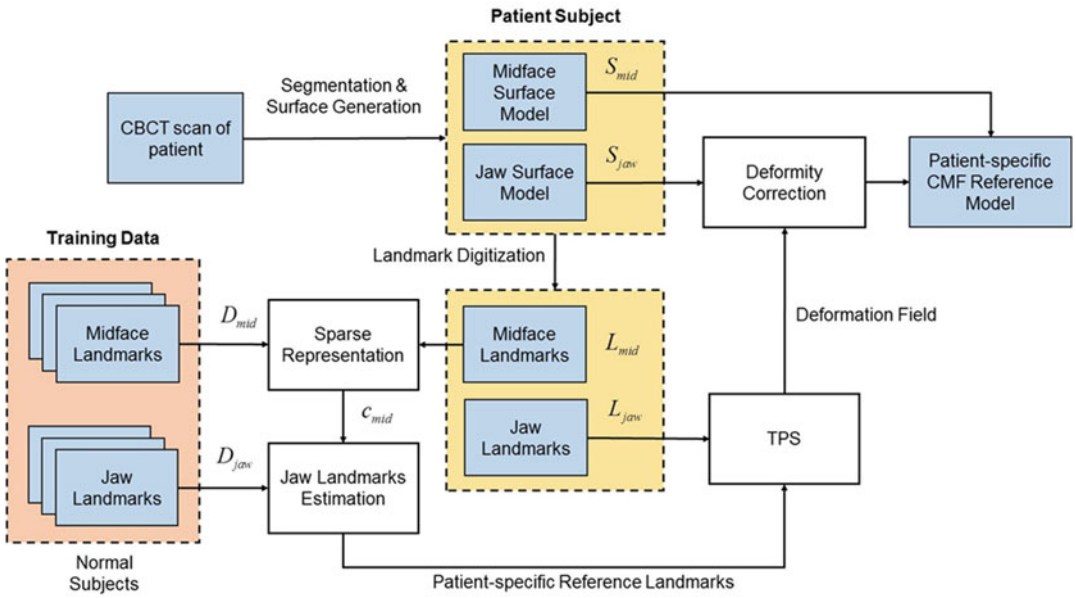


Fig. 9.2 Flowchart of patient-specific jaw model estimation

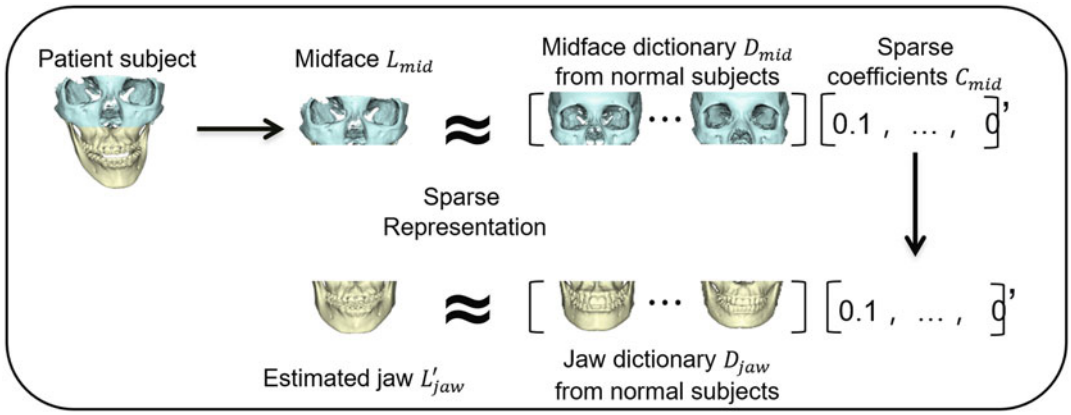


Fig. 9.3 Flowchart of estimating the patient-specific anatomically correct jaw shape. All the shapes in the figure are in terms of landmarks

listed for each landmark. The midface model, S_{mid}, is cyan and the jaw model, S_{jaw}, is yellow. The landmarks on the midface model, S_{mid}, are represented by landmarks L_{mid} and those on the jaw model, S_{jaw}, are represented by landmarks L_{jaw}. Here L_{mid} and L_{jaw} are 27*3 and 31*3 matrices, respectively, consisting of the landmark coordinates. Then each landmark coordinate matrix is concatenated into a column vector.

The flowchart of normal jaw shape estimation is shown in Fig. 9.3. A total of N normal subjects are used as the database. Each subject I^j $(j = 1, \cdots, N)$ is aligned with the patient's midface by linearly aligning the normal subject's midface landmarks, L^j_{mid}, onto the corresponding landmarks, L_{mid}, of the patient I. Let the aligned midface landmarks be $\overset{\sim j}{L}_{\text{mid}}$. After aligning the patient with each normal subject, we can build a dictionary matrix $D_{\text{mid}} = \left[\overset{\sim 1}{L}_{\text{mid}}, \cdots, \overset{\sim N}{L}_{\text{mid}}\right]$. To use D_{mid} to sparsely represent the patient's midface landmarks L_{mid}, we estimate the sparse coefficients, C_{mid}, by solving the optimization problem below:

$$\arg\min_{C_{\text{mid}}} \| D_{\text{mid}} C_{\text{mid}} - L_{\text{mid}} \|^2, \qquad \text{s.t.} \| C_{\text{mid}} \|_0 \leq k \tag{9.1}$$

where $\|\bullet\|$ is the $l2$ norm of vector and $\|\bullet\|_0$ is the $l0$ norm. Here $\|C_{\text{mid}}\|_0$ counts the number of nonzero elements of vector C_{mid}. Parameter k is used to enforce the sparsity of C_{mid}. Intuitively, we are calculating the best possible explanation of L_{mid} as a linear combination of a limited number (less than k) of columns from D_{mid}. While the above problem is NP-hard, it was proven that (1) can be efficiently solved by its $l1$-norm-relaxed version [10, 17]:

$$\arg\min_{C_{\text{mid}}} \| D_{\text{mid}} C_{\text{mid}} - L_{\text{mid}} \|^2 + \lambda \| C_{\text{mid}} \|_1 \tag{9.2}$$

where $\|\bullet\|_1$ is the $l1$ norm of vector, and parameter λ controls the sparsity of representation. Here, $\|D_{\text{mid}} C_{\text{mid}} - L_{\text{mid}}\|^2$ is the data fitting term, and $\lambda \|C_{\text{mid}}\|_1$ is the $l1$ regularization term to enforce the sparsity of coefficients C_{mid}. This problem is equivalent to the well-studied Least Absolute Shrinkage and Selection Operator (LASSO) and can be numerically solved with the method proposed by Tibshirani [18]. This solved sparse coefficient C_{mid} is then applied on the patient's jaw to estimate the normal patient-specific jaw landmarks as:

$$L'_{\text{jaw}} = D_{\text{jaw}} C_{\text{mid}} \tag{9.3}$$

where $D_{\text{jaw}} = \left[\overset{\sim 1}{L}_{\text{jaw}}, \cdots, \overset{\sim N}{L}_{\text{jaw}}\right]$ is the matrix containing all the aligned jaw landmarks from the normal subjects.

We calculate the correspondences between the patient's jaw landmarks L_{jaw} and the estimated reference jaw landmarks L'_{jaw} [19, 20]. Then we interpolate the dense deformation field using thin plate spline (TPS) [17, 21] based on the correspondences. The dense deformation field is then applied to the patient's jaw model S_{jaw}. We derive the patient-specific jaw reference model S'_{jaw}. Finally, we combine the estimated reference model, S'_{jaw}, with the patient's original midface, S_{mid} (which is considered as anatomically normal shape) as the patient-specific CMF reference model.

9.3 Accuracy Evaluation

We have evaluated our proposed method both qualitatively and quantitatively on synthetic and real patient data. The method was implemented with Least-Angle Regression (LARS) algorithm, which was available in the SPArse Modeling Software (SPAMS) toolbox (http://spams-devel.gforge.inria.fr), to solve the LASSO problem. The algorithm was run on a regular office personal computer.

9.3.1 Data Acquisition

We used 30 normal subjects' and 12 patients' data during the evaluation. The 30 sets of Health Insurance Portability and Accountability Act (HIPAA) de-identified multi-slice CT (MSCT) scans of normal subjects were obtained from the Department of Oral and Craniomaxillofacial Surgery at Shanghai 9th People Hospital, Shanghai Jiao Tong University School of Medicine. The data had been collected for a purpose other than this study [22]. All normal data were scanned with a 64-slice GE scanner following a standard clinical protocol: a matrix of 512×512, a field of view of 25 cm, and a slice thickness of 1.25 mm. In addition, we also collected 12 sets of cone-beam CT (CBCT) scans of patients who had already undergone double-jaw orthognathic surgery to correct their jaw deformities. They were randomly selected from clinical archives from the Department of Oral and Maxillofacial Surgery at Houston Methodist Hospital. All patient data were scanned with an i-CAT scanner (Imaging Sciences International LLC, Hatfield, PA) following the standard 0.4 mm isotropic voxel scanning protocol. Institutional review board approval (IRB# 0813-0145) was obtained and all of the subjects were HIPAA de-identified, prior to the study. The data was preprocessed with a CBCT-dedicated method, which segmented and generated the midface and jaw models [23]. The landmarks were manually digitized by experienced oral and maxillofacial surgeons (Z.T. and K.-C. C).

9.3.2 Parameter Setting

In our proposed method, the parameter λ is used to control the sparsity of representation. However, this parameter is unknown at the beginning of the process. Therefore, we used leave-one-out cross-validation of the 30 normal subjects to determine the value of λ [24]. For each normal subject, we first identified its midface landmarks by the landmark dictionary constructed from the other 29 normal subjects. Then we estimated the jaw landmarks by the sparse coefficients obtained from the first step. Last, we measured the distance error between the estimated jaw landmarks and original jaw landmarks. The mean distance errors were obtained with different values of parameter $\lambda \in [0\ 0.1]$. It was observed that the mean distance errors were much smaller, and there was no significant difference when $\lambda = \{1e-5, 1e-4, 1e-3, 0.01\}$. Thus in the following experiments, we chose $\lambda = 0.001$.

9.3.3 Qualitative Evaluation

We evaluated our method with both synthetic and real subjects. In the synthetic evaluation, we created three synthetic 3D models on a randomly selected normal subject to represent the three common types of jaw deformities. These deformities included (a) mandibular hypoplasia with severe anterior open-bite; (b) mandibular hyperplasia; and (c) mandibular asymmetry (vertical

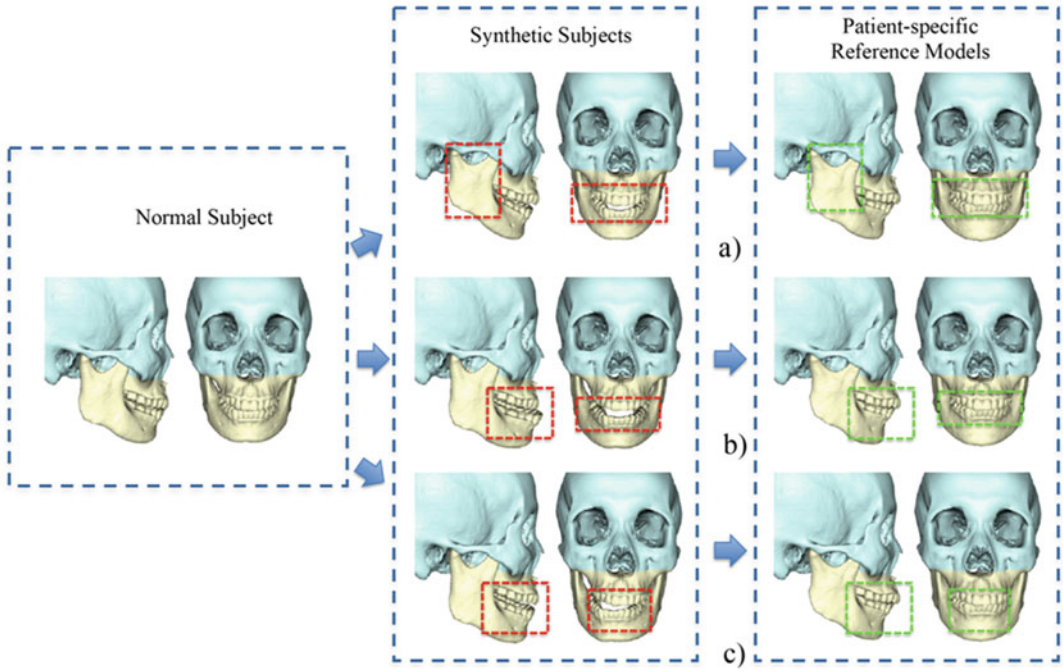


Fig. 9.4 Experiments on synthetic subjects. The left column shows a normal subject, the middle column shows three synthetic patients created from the same subject, with three types of deformities including (**a**) Mandibular hypoplasia, (**b**) Mandibular hyperplasia, and (**c**) Mandibular asymmetry. The right column shows our estimated patient-specific reference models

unilateral condylar hyperplasia) (shown in Fig. 9.4). The morphing process was performed by two experienced oral and maxillofacial surgeons together to ensure that the synthetic deformities were able to reflect the real clinical conditions (Z.T. and K.-C. C). The original model served as the ground truth. Our method estimated a reference model for each type of deformity. During the estimation, the selected normal subject was taken out from the normal data dictionary. The estimation result is shown in Fig. 9.4. All three synthetic deformed models were proposed with a normal shape reference model by our algorithm. All reference models were very similar to the ground truth, which showed that our method was successful with the synthetic models.

In the patient evaluation, 12 real patient models were used. All 30 normal subjects were included in the normal data dictionary. The results are shown in Fig. 9.5. In the figure, each patient has both the original model and the estimated model compared side-by-side. The diagnoses for each patient are also included. The calculated patient-specific reference models were evaluated together by two outside CMF surgeons who were not involved in this project. The result showed that our method was able to estimate the normal shape jaw models for the real patients.

9.3.4 Quantitative Evaluation

We used a new measurement, *normality score*, to quantitatively evaluate the deformity of a subject. The normality score ranges from 0 to 1. A score "1" indicates that the subject is "normal," while score "0" indicates the subject is "patient." The higher the score is, the more likely that the subject is "normal," otherwise, the subject is more likely to be "patient." This score is calculated by sparse representation. Similar with the method described in Sect. 9.2, we built a dictionary with both normal people and patients and their corresponding normality score (either 0 or 1). If a new subject

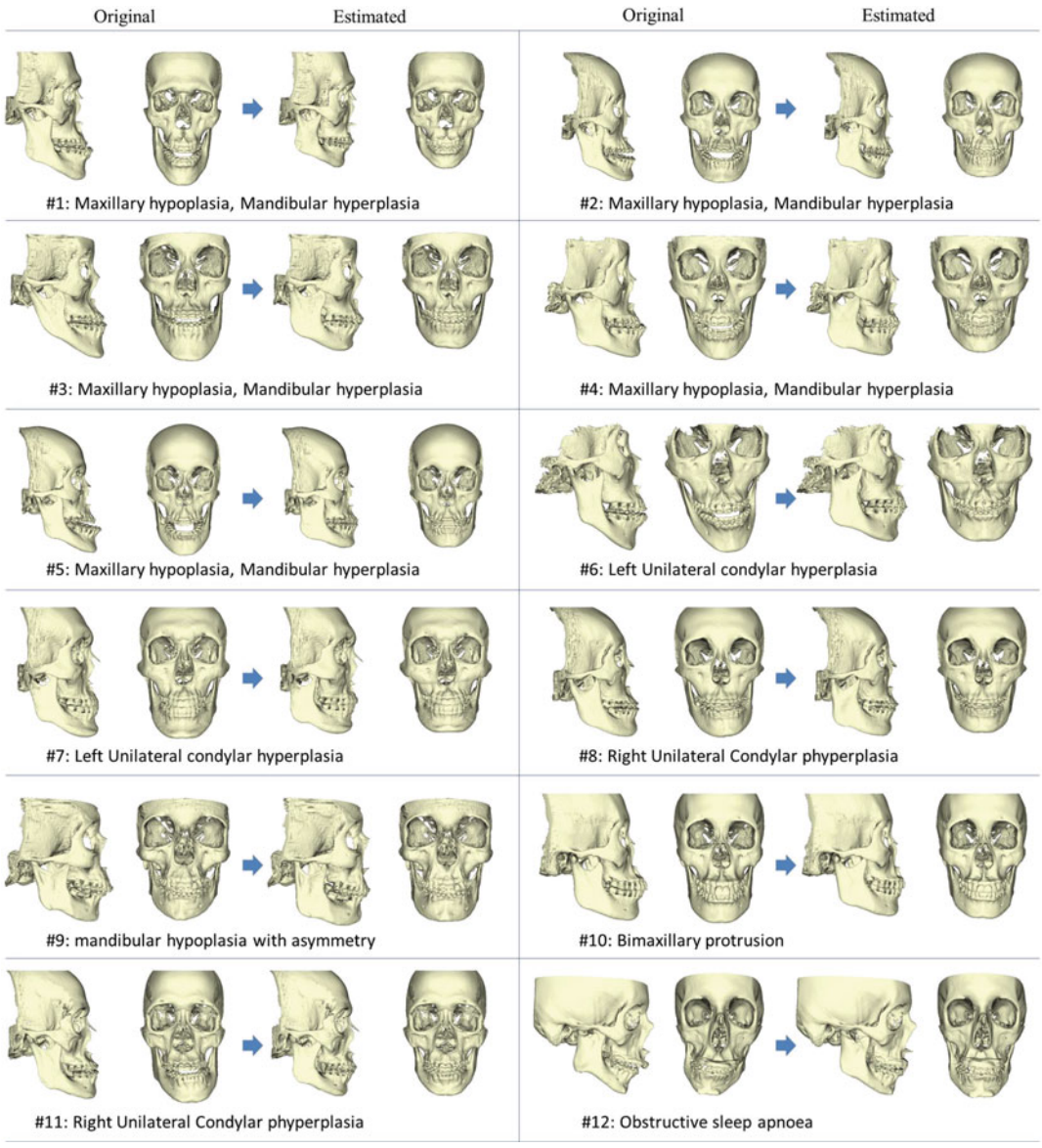


Fig. 9.5 Experiments on real patients. For each patient, the original abnormal CMF is shown in the left and the estimated CMF reference model is shown in the right. The diagnoses of these patients are also shown in the bottom of each subfigure

can be represented by more normal subjects in the dictionary, the subject is more likely to be normal and vice versa. First, we linearly aligned each training subject into a common space, S_{common}, based on the landmarks. The landmarks were concatenated into a column vector. We built a dictionary, D_{all}, by column-wisely stacking the landmark column vectors of all training subjects. Then, we used the dictionary D_{all} to sparsely represent the new testing subject by calculating the sparse coefficient vector C. Vector v is a column vector with the scores of the each training subject. Finally, we calculated the normality score of the new subject as $C^T v$. Intuitively, the normality score was a weighted average of those training subjects which best represented the new subject.

We calculated normality scores of the three groups of subjects, including 30 normal subjects,

Table 9.1 The mean and standard deviation of the normality scores for the Normal subjects, the Patients, the corresponding reference models by our proposed method and majority voting

Group	Normal	Patient	Proposed reference	Majority voting
Normality score	0.787 ± 0.153	0.328 ± 0.273	0.823 ± 0.085	0.7624 ± 0.087

12 patients with jaw deformities, and the reconstructed patient-specific reference models. We also the calculated normality score of the reference models reconstructed by a majority voting method as a comparison. The majority voting method simply took the average of all the jaw landmarks from the aligned normal subjects. The score calculation was done in a leave-one-out manner. The testing subject was excluded from the dictionary construction. The mean and standard deviation are summarized in Table 9.1. The result showed that typical normal subjects had normality scores greater than 0.5, while typical patients had normality scores less than 0.5. This finding indicates that the proposed normality scores can effectively distinguish between normal subjects and patients. The normality score of the patient-specific models reconstructed by our method showed that they were more likely to be normal. These scores were significantly higher than the original patient models (p-value<0.01). Thus, our proposed method was able to effectively reconstruct normal-shape patient-specific reference models. In addition, the normality score of the reconstructed models was slightly higher than that of the real normal subjects. This finding was due to the fact that human faces are not perfectly symmetrical. There is about 5% of fluctuating asymmetry in the normal population [25]. Also, the upper dental midline might have a deviation as large as 2 mm in some normal subjects. However, the reference models reconstructed by our method could be perfectly symmetric with the upper dental midline well-aligned with the midsagittal plane. The normality score of the models reconstructed by majority voting was smaller than both the normal subjects and the models reconstructed by our method. The comparison showed that our method was superior to the majority voting method in reconstructing a patient-specific CMF normal shape model.

9.4 Discussion

In this chapter, we presented a novel method to reconstruct a patient-specific reference model for treating patients with jaw deformities. We use the spare representation technique and models with normal CMF shape to estimate an anatomically correct reference model for a patient who has a normal cranium and midface but deformed maxilla and mandible. The estimated reference model is correlated with the patient's cranium and midface and thus it is more suitable for personalized patient care than using the averaged cephalometric values. The orthognathic surgery can restore the deformed jaws to normal shape by matching the bony segments to the reference model with rigid transformation [26]. This is an important step toward personalized medicine.

Although we tested our method on the patients who had jaw deformities, the use of our reconstructed models can also be extended to the patients with jaw trauma, in which the posttraumatic reconstructive surgery can remodel the shattered bony segments to the patient-specific reference model (Please see details in Chap. 4).

Acknowledgments This work was supported in part by National Institutes of Health/National Institute of Dental and Craniofacial Research grants DE022676, DE021863 and DE027251. Dr. Chen was sponsored by the Taiwan Ministry of Education, and Dr. Tang was sponsored by the China Scholarship Council while they were working at the Surgical Planning Laboratory, Department of Oral and Maxillofacial Surgery, Houston Methodist Research Institute, Houston, TX, USA.

References

1. Lew TA, Walker JA, Wenke JC, Blackbourne LH, Hale RG. Characterization of craniomaxillofacial battle injuries sustained by United States service members in the current conflicts of Iraq and Afghanistan. J Oral Maxillofac Surg. 2010;68(1):3–7. https://doi.org/10.1016/j.joms.2009.06.006.

2. Xia JJ, Gateno J, Teichgraeber JF. New clinical protocol to evaluate craniomaxillofacial deformity and plan surgical correction. J Oral Maxillofac Surg. 2009;67(10):2093–106. https://doi.org/10.1016/j.joms.2009.04.057.
3. Gateno J, Xia JJ, Teichgraeber JF, Christensen AM, Lemoine JJ, Liebschner MA, Gliddon MJ, Briggs ME. Clinical feasibility of computer-aided surgical simulation (CASS) in the treatment of complex cranio-maxillofacial deformities. J Oral Maxillofac Surg. 2007;65(4):728–34. https://doi.org/10.1016/j.joms.2006.04.001.
4. Swennen GR, Barth EL, Eulzer C, Schutyser F. The use of a new 3D splint and double CT scan procedure to obtain an accurate anatomic virtual augmented model of the skull. Int J Oral Maxillofac Surg. 2007;36(2):146–52. https://doi.org/10.1016/j.ijom.2006.09.019.
5. Swennen GR, Mommaerts MY, Abeloos J, De Clercq C, Lamoral P, Neyt N, Casselman J, Schutyser F. The use of a wax bite wafer and a double computed tomography scan procedure to obtain a three-dimensional augmented virtual skull model. J Craniofac Surg. 2007;18(3):533–9. https://doi.org/10.1097/scs.0b013e31805343df.
6. Xia J, Ip HH, Samman N, Wang D, Kot CS, Yeung RW, Tideman H. Computer-assisted three-dimensional surgical planning and simulation: 3D virtual osteotomy. Int J Oral Maxillofac Surg. 2000;29(1):11–7.
7. Zhang S, Zhan Y, Dewan M, Huang J, Metaxas DN, Zhou XS. Towards robust and effective shape modeling: sparse shape composition. Med Image Anal. 2012;16(1):265–77. https://doi.org/10.1016/j.media.2011.08.004.
8. Wang G, Zhang S, Li F, Gu L. A new segmentation framework based on sparse shape composition in liver surgery planning system. Med Phys. 2013;40(5):051913. https://doi.org/10.1118/1.4802215.
9. Zhang S, Zhan Y, Metaxas DN. Deformable segmentation via sparse representation and dictionary learning. Med Image Anal. 2012;16(7):1385–96. https://doi.org/10.1016/j.media.2012.07.007.
10. Donoho D. For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution. Commun Pure Appl Math. 2006;59:797–829.
11. Vannier MW, Marsh JL, Warren JO. Three dimensional CT reconstruction images for craniofacial surgical planning and evaluation. Radiology. 1984;150(1):179–84. https://doi.org/10.1148/radiology.150.1.6689758.
12. Xia JJ, Gateno J, Teichgraeber JF. Three-dimensional computer-aided surgical simulation for maxillofacial surgery. Atlas Oral Maxillofac Surg Clin North Am. 2005;13(1):25–39. https://doi.org/10.1016/j.cxom.2004.10.004.
13. Zachow S, Lamecker H, Elsholtz B, Stiller M. Reconstruction of mandibular dysplasia using a statistical 3D shape model. Int Congr Ser. 2005;1281 https://doi.org/10.1016/j.ics.2005.03.339.
14. Zhu Y, Papademetris X, Sinusas AJ, Duncan JS. Segmentation of the left ventricle from cardiac MR images using a subject-specific dynamical model. IEEE Trans Med Imaging. 2010;29(3):669–87. https://doi.org/10.1109/TMI.2009.2031063.
15. Zhang W, Yan P, Li X. Estimating patient-specific shape prior for medical image segmentation. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 30 March–2 April 2011, 2011. pp 1451–1454. https://doi.org/10.1109/ISBI.2011.5872673
16. Xia JJ, McGrory JK, Gateno J, Teichgraeber JF, Dawson BC, Kennedy KA, Lasky RE, English JD, Kau CH, McGrory KR. A new method to orient 3-dimensional computed tomography models to the natural head position: a clinical feasibility study. J Oral Maxillofac Surg. 2011;69(3):584–91. https://doi.org/10.1016/j.joms.2010.10.034.
17. Starck JL, Elad M, Donoho DL. Image decomposition via the combination of sparse representations and a variational approach. IEEE Trans Image Process. 2005;14(10):1570–82. https://doi.org/10.1109/tip.2005.852206.
18. Tibshirani R. Regression Shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88.
19. Lapeer RJ, Prager RW. 3D shape recovery of a newborn skull using thin-plate splines. Comput Med Imaging Graph. 2000;24(3):193–204. https://doi.org/10.1016/s0895-6111(00)00019-7.
20. Xue Z, Shen D, Davatzikos C. Statistical representation of high-dimensional deformation fields with application to statistically constrained 3D warping. Med Image Anal. 2006;10(5):740–51. https://doi.org/10.1016/j.media.2006.06.007.
21. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell. 2009;31(2):210–27. https://doi.org/10.1109/TPAMI.2008.79.
22. Yan J, Shen G-f, Fang B, Shi H-m, Wu Y, Shao Z-y, Xia B-q, Yu D-d. Three-dimensional CT measurement for the craniomaxillofacial structure of normal occlusion adults in Jiangsu, Zhejiang and Shanghai Area. China J Oral Maxillof Surg. 2010.
23. Wang L, Chen KC, Gao Y, Shi F, Liao S, Li G, Shen SG, Yan J, Lee PK, Chow B, Liu NX, Xia JJ, Shen D. Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization. Med Phys. 2014;41(4):043503. https://doi.org/10.1118/1.4868455.
24. Mairal J, Bach F, Ponce J. Task-driven dictionary learning. IEEE Trans Pattern Anal Mach Intell. 2012;34(4):791–804. https://doi.org/10.1109/TPAMI.2011.156.

25. Gateno J, Jones TL, Shen SGF, Chen KC, Jajoo A, Kuang T, English JD, Nicol M, Teichgraeber JF, Xia JJ. Fluctuating asymmetry of the normal facial skeleton. Int J Oral Maxillofac Surg. 2018;47(4):534–40. https://doi.org/10.1016/j.ijom.2017.10.011.

26. Wu G, Jia H, Wang Q, Shen D. SharpMean: groupwise registration guided by sharp mean image and tree-based registration. NeuroImage. 2011;56(4):1968–81. https://doi.org/10.1016/j.neuroimage.2011.03.050.

Part III

Machine Learning and Dental Designs

Machine (Deep) Learning for Orthodontic CAD/CAM Technologies

10

Tai-Hsien Wu, Chunfeng Lian, Christian Piers, Matthew Pastewait, Li Wang, Dinggang Shen, and Ching-Chang Ko

10.1 Introduction

Many modern orthodontic clinics around the world use computer-aided design and computer-aided manufacturing (CAD/CAM) technologies for diagnosis and treatment planning. The three-dimensional (3D) dental surface mesh model—a 3D mesh model of a single dental arch—is the most well-known example of CAD/CAM in orthodontics and has been widely used to help dentists and orthodontists extract, move, and rearrange teeth for simulation of treatment outcomes, as well as for aligner design. Digital scanning has been extensively utilized for acquiring 3D dental mesh models. Intraoral scanning has largely replaced traditional impressions for the construction of the digital surface model. Compared with conventional impressions, intraoral scanning is more time efficient and more comfortable for patients. Additionally, intraoral scanning reduces the risk of allergies caused by many constituents of physical impression materials [1].

Accurate tooth segmentation and partition from the digitized dental mesh model is a fundamental step in most of the orthodontic applications related to dental mesh models. However, this task is still challenging. The following major difficulties exist [2, 3]:

- The shape of teeth varies dramatically between different patients.
- Patients' teeth are often positioned abnormally. For example, neighboring teeth that are

T.-H. Wu (✉) · C.-C. Ko
Division of Orthodontics, College of Dentistry, The Ohio State University, Columbus, OH, USA
e-mail: wu.5084@osu.edu; tai-hsien.wu@wmich.edu

C. Lian
School of Mathematics and Statistics, Xi'an Jiaotong University, Shaanxi, China

C. Piers
Private Practice of Orthodontics, Morganton, NC, USA

M. Pastewait
United States Air Force, Kadena AB, Okinawa, Japan

L. Wang
Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC , USA

D. Shen
School of Biomedical Engineering, ShanghaiTech University, Shanghai, China

Department of Research and Development, Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_10

crowded and misaligned can obscure normal interstices.
- Missing/decayed teeth and chips or fractures in teeth are common.
- Deep intraoral regions (e.g., the 2nd/3rd molars) may not be perfectly captured during intraoral scanning.
- Artifact noise can be generated during the scanning (e.g., air bubbles in plaster models).
- Some patients wear dental accessories such as orthodontic brackets on their teeth.

Although several commercial products (e.g., 3Shape) are available to segment the teeth in digitized dental mesh models, they are often labor intensive, and their results are significantly influenced by user operation. Hence, many studies have focused on developing an accurate automatic segmentation method. Despite numerous general automatic mesh segmentation methods existing in computer vision communities, their performance is not satisfactory when applied to dental mesh models due to the aforementioned challenges. Therefore, Liao et al. [2] developed an automatic tooth segmentation based on the harmonic field. They demonstrated that this method is able to partition teeth automatically, efficiently, and with high quality and robustness. Learning-based methods have been comprehensively studied in both computer vision and computer graphics communities, and are also potentially applicable to the task of tooth segmentation, even for patients wearing dental accessories such as orthodontic brackets.

Deep learning, a type of machine learning that includes deep neural networks, has been applied in a variety of fields. In particular, a specific network structure called the convolutional neural network (CNN) has shown outstanding image analysis abilities. These abilities include classification, object detection, and segmentation. While CNNs have achieved great success with both 2D and 3D images, all of the input images are composed of Euclidean data. These are uniform grid structures. The mesh model structure is constituted by a group of triangles (i.e., a group of vertexes and edges), known as non-Euclidean data, to which a CNN network cannot be directly applied. However, a subfield of deep learning called "geometric deep learning" has recently emerged to attempt to generalize deep neural networks to non-Euclidean domains [4].

Simply speaking, there are several paths to achieve this goal. The first method is *voxelization*, where the non-Euclidean data is converted to Euclidean structure, and then the classical CNN can be used on the converted data [5–7]. However, if the voxel spacing is too large, the conversion leads to the loss of natural features (e.g., blurring images). In contrast, if the voxel spacing is too small, the entire process becomes inefficient in terms of computing loading and storage. The second path is the *graph convolutional networks* (GCNs) [8–10]. This technique is inspired by the success of CNNs in the computer vision field. Researchers redefined the notation of convolution for graph data based on spectral and spatial perspectives. Since there has been increasing interest in the GCNs or graph neural networks (GNN), several papers reviewed and summarized the existing graph-based networks and their achievements in this field [4, 11, 12]. However, not all algorithms are suitable for the mesh structure because mesh is only a type of graph structure. For example, the spectral GCN learned on one mesh (or manifold) cannot be transferred to another mesh. This is the major limitation of spectral approaches [13]. In the dental field [3], feature-based deep neural networks (DNNs) are used to predefine hand-crafted geometric features from the 3D data and reshape them as an image so that a multi-stage CNN can be trained for labeling dental mesh surface models. However, such applications of CNNs may lead to unstable segmentations, because they ignore the fact that the input geometric data are actually unordered (i.e., different organizations of them result in different "images"). This implies that a more robust method is still needed.

Recently, a pioneering work, called PointNet, was proposed for end-to-end 3D shape analysis [14]. Directly using raw geometric data (e.g., the coordinates and normal vectors of point clouds) as input, PointNet learns translation-invariant

deep features for shape classification/segmentation, yielding state-of-the-art performance in terms of efficiency and accuracy. Ko et al. demonstrated the segmentation results based on PointNet [15], and it was obvious that the results were not accurate enough because PointNet ignored the local geometric context. However, effectively modeling local structures is critical for the success of deep neural networks in fine-grained segmentation tasks. Since then, some efforts were proposed to extend PointNet by including contextual information [16, 17]. These efforts coarsely group points into several clusters according to their spatial relationship. However, such coarse operations cannot perform well with fine-grained tooth segmentation, especially considering that each tooth only comprises a very small portion of the entire dental surface model.

To improve the major limitation of PointNet, Lian et al. [18, 19] proposed MeshSegNet which is an end-to-end deep neural network specifically for automatic tooth segmentation. MeshSegNet is an extension of PointNet and learns high-level geometric features from the raw dental mesh data. In this chapter, we briefly introduce deep learning and then comprehensively review MeshSegNet and how to use it for the task of automated tooth segmentation, from preprocessing the dental mesh model to post-processing for refining of the segmentation results.

10.2 Background Knowledge in Deep Learning

When we train a deep learning model in supervised learning, what we are really doing is optimizing the parameters of a nonlinear model. A well-trained model captures the generalized trend from given data and will be able to predict the corresponding result of the unseen input. There are several professional terms frequently used in deep learning. Before we go further, let us briefly explain the most important terms so that those readers who are not familiar with deep learning can understand the following sections.

10.2.1 Loss (Cost) Function and Metrics

Deep learning in its simplest form is an optimization problem for finding the minimum of a function called the loss (cost) function. The loss function reduces various good and bad characteristics of a possibly complex system down to a single number, a scalar value. This scalar value allows for discrimination between model results. Simply speaking, the model with the smallest loss function value has the best performance. A "good" loss function must capture all the properties of the problem, so selecting a loss function can be challenging. Another index to assist in the ranking and comparison of models is called metric. A metric is very similar to the loss function, which evaluates the quality of a model by comparing the predicted result from the model with the ground-truth. A key distinction between a loss function and metrics is that metrics have no role in the optimization process. That is, they have no influence in the optimization of training loss functions. Also, a model can have multiple metrics but only has one loss function (or a hybrid of multiple loss functions).

10.2.2 Training, Validation, and Test Datasets

When building models for a supervised machine learning project, the data (i.e., raw data and their labels) must be divided into three subsections of datasets, i.e., training, validation, and testing sets. Training data is used to learn machine learning models. The size of the training dataset depends on the specific project and the diversity of the samples. The validation dataset periodically evaluates the model to ensure that the learning of the model is generalized across the dataset and not overfitted to the training data. (If "overfitted," the model only predicts precisely in the training dataset and does not generalize well to the unseen data.) Both training and validation datasets are involved in the training process. The test dataset

is the data that is used to evaluate the trained model. The test dataset is kept separate during the training process and is only introduced to test the unbiased performance of the model on unseen data.

10.2.3 Epoch and Batch Size

The training dataset is often too large to feed into the network in one shot, so the training data is broken down into batches. Batches can be defined as evenly spliced portions of the training data. An epoch means when each of the batches (i.e., the entire set of training dataset) has been passed through the network one time. For example, if you have a dataset with 1000 samples, it can be divided into 10 batches with 100 samples in each batch. You must pass all 10 batches through the neural network to complete a single epoch.

10.3 Automatic Tooth Segmentation Based on MeshSegNet

The raw data of a dental model is a surface mesh structure, which is constituted by a group of cells (i.e., triangles). Each cell is formed by three points and three edges, and the adjacent cells share the contacted points and edges. To segment the dental surface model, we suggest labeling each individual cell instead of each point because a point might be involved in multiple cells with different categories. In order to express the detail of teeth, the dental model usually uses approximately 100,000 cells to represent the entire surface of the dental model—which is, again, a single dental arch. The size of data is sometimes too large to store and access, particularly in Graphics Processing Unit (GPU) memory. Therefore, simplification or so-called "down-sampling" is a required preprocess step of automatic tooth segmentation.

10.3.1 PreProcessing

10.3.1.1 Simplification

Mesh simplification is necessary for the preprocessing of dental meshes to boost computing efficiency. A good simplification method must preserve not only the entire outline but also the detailed features. Xu et al. proposed a boundary-ware tooth simplification [3]. Here, we simply introduce an open-source toolkit, the *Visualization Toolkit* (VTK), which includes a feature-preserving mesh simplification function *vtkQuadricDecimation*. For the complete implementation of this function, please refer to the VTK textbook [20]. Figure 10.1 is an example, where (a) represents an original dental mesh model from an iTero® Element scanner (http://www.itero.com) and (b) represents a simplified model using the VTK function, respectively.

10.3.1.2 Data Annotation

In supervised learning, we need to annotate labels on data and make a machine learning model to learn the relationship between data and labels. In our task, the data are dental meshes and la-

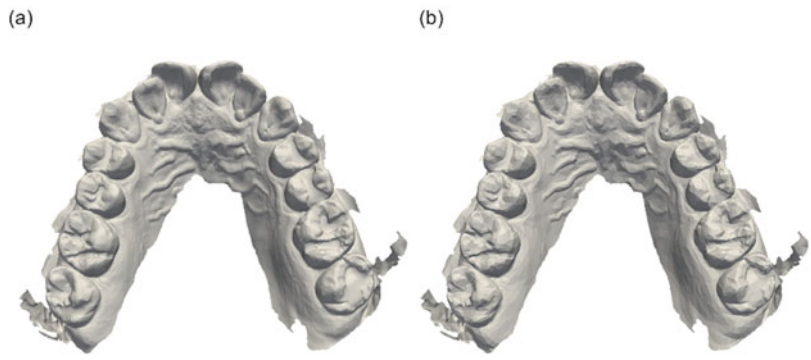


Fig. 10.1 (**a**) An original dental mesh model from the iTero® Element scanner, which consists of 100,000 cells and 51,255 points. (**b**) A simplified dental mesh using VTK, which consists of 9999 cells and 5378 points

bels are the manual segmentation of the dental meshes. The original dental models are usually in Stereolithography (STL) format, and there is no particular software that can assign a label on a single cell (i.e., triangle) efficiently. Therefore, we develop a graphical user interface (GUI) program, called *Mesh Labeler*, which can annotate a dental arch model easily. To download and learn how to use the program, please refer to its website https://github.com/Tai-Hsien/Mesh_Labeler.

10.3.1.3 Data Augmentation

Deep learning requires a large amount of data to obtain excellent performance. The data augmentation technique is a common way to increase the amount of data when the number of samples is insufficient. In [18, 19], they are augmented by two operations. First, the authors randomly rotated, translated, and rescaled each 3D dental mesh in the training and validation datasets. Second, they randomly sampled 50% of the cells from 14 teeth (from the upper left 7 to upper right 7, each tooth 3.57%) and 50% of the cells from the gingiva as the network input (with 6000 cells in total). This means the input samples were "different" in every epoch during the training process. The combination of the two operations could largely enrich the training set and also mitigate the imbalanced learning challenge, caused by the fact that each tooth comprises only a very small part of the whole dental surface. An augmented dataset can improve the generalization ability of trained networks.

10.3.2 MeshSegNet

The network architecture of MeshSegNet is shown in Fig. 10.2. For those seeking an in-depth description of all operations and their functions in MeshSegNet, from input layer to output layer, the following section will be helpful. For those seeking an overview comparing PointNet to MeshSegNet, proceed to the "Segmentation Result Comparison" section.

In the input layer, the raw dental surface mesh generates two types of input data: Mesh data $\boldsymbol{F}^0$ and adjacency matrices $\boldsymbol{A}_S$ and $\boldsymbol{A}_L$. The input mesh data $\boldsymbol{F}^0$ with the size of $N \times 15$ is "partial data" of the raw mesh surface data (see **Data Augmentation**), where N is the number of sampled mesh cells (triangles). Each cell in $\boldsymbol{F}^0$ is initially described by a 15-dimensional row vector, which are the coordinates of the three vertices (9 units, i.e., X, Y, and Z components for each vertex), the normal vector (3 units, i.e., X, Y, and Z components), and the relative position (3 units, i.e., X, Y, and Z components) with respect to the whole surface (i.e., the centroid of whole mesh). The input adjacency matrices $\boldsymbol{A}_S$ and $\boldsymbol{A}_L$ both are $N \times N$ adjacency matrices, which are computed based on two neighborhood balls with different radiuses (subscripts S and L denote small and large, respectively). If a distance between two centroids of cells is smaller than the neighborhood ball, the corresponding element in the adjacency matrix is 1. Otherwise, the element is 0. An example of adjacency matrix is shown in Fig. 10.3a. In MeshSegNet, the adjacency matrices provide the graph connections between any two cells in the underlying Euclidean space.

The MeshSegNet consists of three multi-layer perceptrons [21] (MLPs, i.e., MLP-1, MLP-2, and MLP-3), one feature transformer module (FTM), two graph-constrained learning modules (GLMs, i.e., GLM-1 and GLM2), and a final one-dimensional (1D) convolutional layer (Conv), as shown in Fig. 10.2. MLP-1 contains two 1D Convs, both with 64 channels. MLP-2 has three 1D Convs, each with 64, 128, and 512 channels, respectively. MLP-3 contains four 1D Convs, each with 256, 256, 128, and 128 channels, respectively. Denote $\boldsymbol{F}^1 \in \mathbb{R}^{N \times 64}$ as the features learned by MLP-1. The FTM predicts a 64×64 transformation matrix $\boldsymbol{T}$ from $\boldsymbol{F}^1$, and directly updates the feature matrix as $\hat{\boldsymbol{F}}^1 = \boldsymbol{F}^1\boldsymbol{T}$. To learn the 64×64 feature transformation matrix $\boldsymbol{T}$, FTM employs six 1D Convs with 64, 128, 512, 256, 128, and 64^2 channels, respectively, where each of the first five layers is followed by the batch normalization (BN) and rectified linear unit (ReLU) activation function, while the last layer (without BN and ReLU) is followed by a tensor reshape operation.

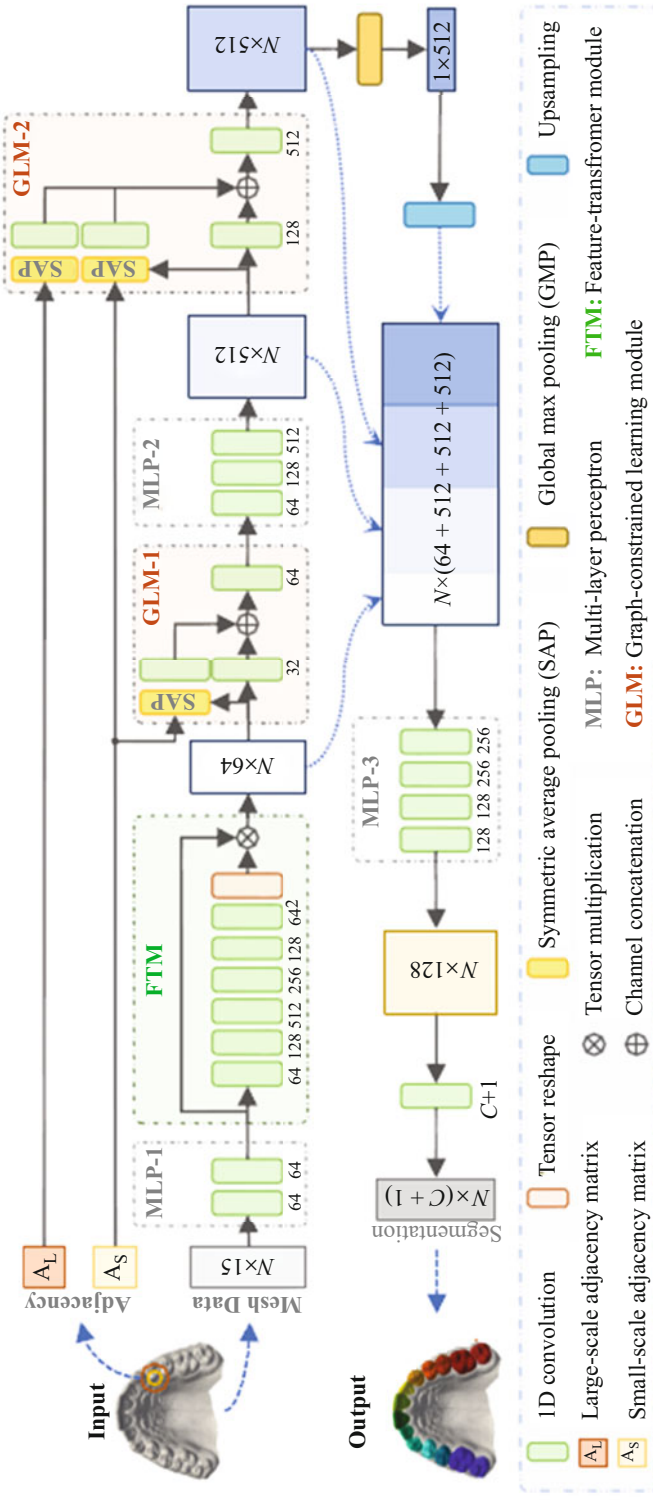


Fig. 10.2 Illustration of MeshSegNet Architecture, a multi-scale deep neural network to learn high-level geometric features for end-to-end tooth segmentation on 3D dental surfaces. Reprinted from Ref. [18] with permission the Springer, Cham

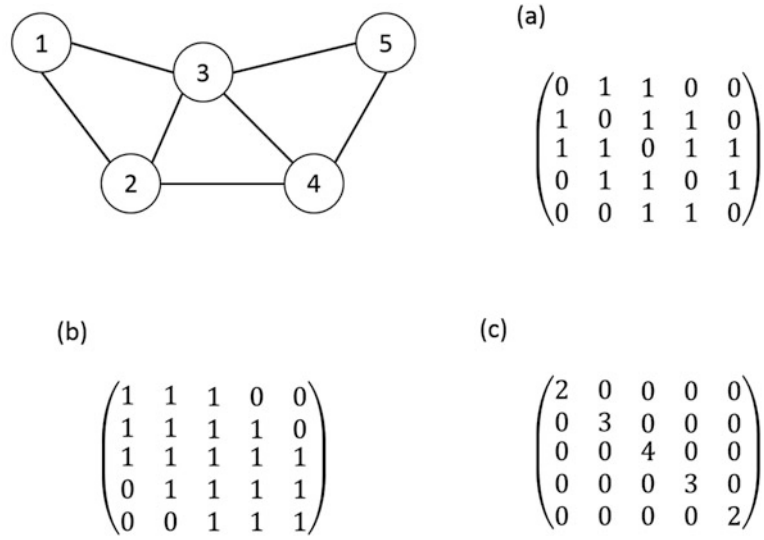


Fig. 10.3 Examples of a mesh and its (**a**) adjacency matrix $\boldsymbol{A}$, (**b**) adjacency matrix with self-loop $\tilde{\boldsymbol{A}}$, and (**c**) diagonal degree matrix $\boldsymbol{D}$

There are two major innovations in MeshSegNet that led to its enhanced performance, compared to the original PointNet. The first innovation is the *multi-scale graph constrained learning* which is implemented by GLMs. In GLM-1, a graph-based fusion operation (called symmetric average pooling, SAP) is first applied on the output of FTM ($\hat{\boldsymbol{F}}^1 \in \mathbb{R}^{N\times 64}$) to propagate the contextual information (provided by neighboring cells) onto each centroid cell. The resulting feature matrix $\tilde{\boldsymbol{F}}^1 \in \mathbb{R}^{N\times 64}$ encoding local geometric context has the form as follows:

$$\tilde{\boldsymbol{F}}^1 = \left(\tilde{\boldsymbol{D}}_S^{-\frac{1}{2}}\tilde{\boldsymbol{A}}_S\tilde{\boldsymbol{D}}_S^{-\frac{1}{2}}\right)\hat{\boldsymbol{F}}^1 \sim \left(\tilde{\boldsymbol{D}}_S^{-1}\tilde{\boldsymbol{A}}_S\right)\hat{\boldsymbol{F}}^1, \tag{10.1}$$

where $\tilde{\boldsymbol{A}}_S = \boldsymbol{A}_S + 1$ is regarded as an adjacency with self-loops, $\tilde{\boldsymbol{D}}_S^{-\frac{1}{2}}\tilde{\boldsymbol{A}}_S\tilde{\boldsymbol{D}}_S^{-\frac{1}{2}}$ is the respective symmetric-normalized adjacency, and $\tilde{\boldsymbol{D}}_S$ is the diagonal degree matrix. Figure 10.3b and c also illustrates what an adjacency matrix with self-loop and diagonal degree matrix are, respectively. In practical implementation, $\tilde{\boldsymbol{D}}_S^{-1}\tilde{\boldsymbol{A}}_S$ is used to replace $\tilde{\boldsymbol{D}}_S^{-\frac{1}{2}}\tilde{\boldsymbol{A}}_S\tilde{\boldsymbol{D}}_S^{-\frac{1}{2}}$ for reducing computation. After SAP in GLM-1, both $\tilde{\boldsymbol{F}}^1$ and $\hat{\boldsymbol{F}}^1$ are further squeezed by shared-weights 1D Conv with 32 channels. The resultant feature matrices are then concatenated across channels, followed by the fusion by another 1D conv with 64 channels. Different from GLM-1, GLM-2 enlarges the receptive field and learns multi-scale contextual features. Specifically, similar to Eq. (10.1), the $N \times 512$ feature matrix from MLP-2 (called $\boldsymbol{F}^2 \in \mathbb{R}^{N\times 512}$) is processed by two parallel SAPs in terms of $\boldsymbol{A}_S$ and $\boldsymbol{A}_L$, respectively. The resultant feature matrices and $\boldsymbol{F}^2$ are then squeezed by shared-weights 1D Convs with 128 channels which are finally concatenated across channels and fused by another 1D Conv with 512 channels. All 1D Convs in these MLPs and the GLMs are followed by the BN and ReLU.

The second innovation is the ***dense fusion of local-to-global features***. Similar to PointNet, a global max pooling (GMP) is applied on the output of GLM-2 to produce the translation invariant holistic features, aiming to encode the semantic information of the whole dental mesh surface. Different from PointNet which inserts only skip connections between cell/point-wise and holistic features, MeshSegNet assumes that the multi-scale contextual features (produced by intermediate GLMs) could provide additional information to comprehensively describe mesh cells. Therefore, MeshSegNet densely concatenates the local-to-global features from FTM, GLM-1, GLM-2,

and upsampled GMP, followed by MLP-3 to yield a $N \times 128$ feature matrix. Based on this feature matrix, an output layer (1D Conv with softmax activation) is used to predict a $N \times (N_T + 1)$ probability matrix, where N_T and 1 represent the number of teeth ($N_T = 14$ in [18, 19]) and gingiva. In the probability matrix, each row denotes the probabilities of the respective cell belonging to specific categories. The two innovations in MeshSegNet were comprehensively verified in [18, 19], proving that they significantly enhance the performance in the mesh segmentation process.

The implementation here was the same as in [18, 19], carried out by using Python based on Keras (a high-level neural networks application programming interface) and trained by minimizing the generalized Dice loss [22] using the AMSGrad variant [23] of the Adam optimizer with a batch size of 10 and 200 epochs. A threefold cross-validation was performed in this task. The raw dataset consists of 20 maxillary dental mesh models from different patients acquired using iTero® Element scans. The original dental surfaces roughly contain 100,000 mesh cells, which were simplified to 10,000 cells, as described in **Simplification**. The ground-truth segmentations for 14 teeth were manually annotated on simplified surface meshes.

10.3.3 Segmentation Result

Three metrics were used to evaluate the segmentation results, including Dice similarity coefficient (DSC), sensitivity (SEN), and positive prediction value (PPV). The three metrics can be explained using a confusion matrix (see Table 10.1), where predicted positive and negative represent the predicted labels based on the machine learning system, while ground-truth positive and negative represent the true labels. The DSC can be expressed as follows:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TP} + \mathrm{FN})}, \tag{10.2}$$

where TP, FP, and FN represent True Positive, False Positive, False Negative, as given in Table 10.1. Moreover, the SEN and PPV are another two measures used widely in image segmentation, as follows:

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{10.3}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{10.4}$$

The DSC indicates the similarity between the predicted results and ground-truth, while the SEN and PPV reveal the probability that the prediction is positive based on all positives in ground-truth and in prediction, respectively. All these three metrics have ranges between 0 and 1. A higher value can be said to correspond to higher prediction accuracy.

Figure 10.4 shows the comparison of segmentation results among PointNet, MeshSegNet, and ground-truth. In addition, the results of overall segmentation for all teeth and individual teeth in terms of the three metrics are given in Table 10.2 and Fig. 10.5.

10.3.3.1 Segmentation Result Comparison

Compared with the state-of-the-art PointNet method, MeshSegNet led to significantly better results, suggesting that MeshSegNet could effectively capture and leverage local geometric context to improve the segmentation performance. As shown in Fig. 10.4 and Table 10.2, the MeshSegNet method has an overall better performance. A paired t-test was carried out to investigate the improvement

Table 10.1 Confusion matrix

	Ground-Truth Positive	Ground-Truth Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

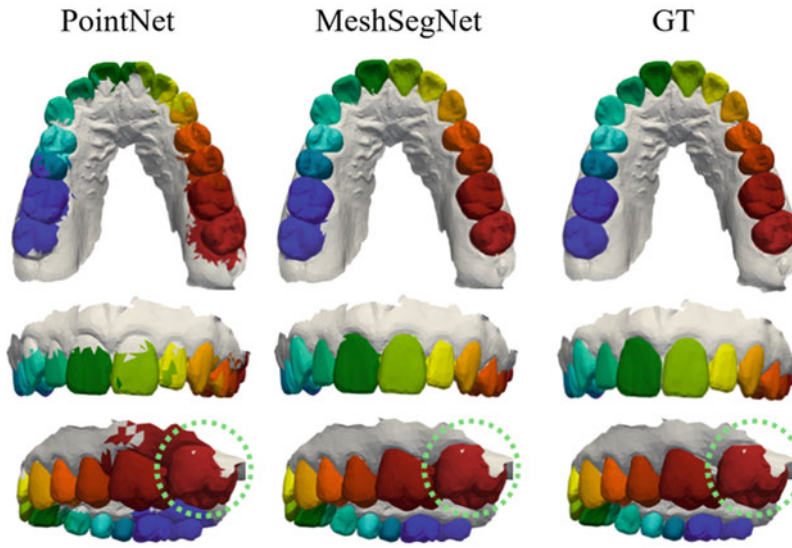


Fig. 10.4 Comparison of segmentation results among PointNet, MeshSegNet, and ground-truth (GT). Compared -with PointNet, MeshSegNet effectively reduced false positives for segmenting molars and false negatives for segmenting central incisors, as shown in the first and second rows. Particularly, MeshSegNet annotated molars more precisely even though molars were not completely captured by IOS, as shown in green circles in the third row. Reprinted from Ref. [18] with permission from Springer, Cham

Table 10.2 Segmentation results (mean ± standard deviation) for all teeth quantified in terms of Dice similarity coefficient (DSC), sensitivity (SEN), and positive prediction value (PPV), under threefold cross-validation, where p indicates the p-value for the statistical significance comparison between MeshSegNet and PointNet

Metric	PointNet	MeshSegNet
DSC	0.781 ± 0.134 $p = 1.6 \times 10^{-10}$	0.938 ± 0.060 n/a
SEN	0.828 ± 0.167 $p = 8.1 \times 10^{-7}$	0.946 ± 0.062 n/a
PPV	0.766 ± 0.163 $p = 6.1 \times 10^{-9}$	0.934 ± 0.077 n/a

Reprinted from Ref. [18] with permission from Springer, Cham

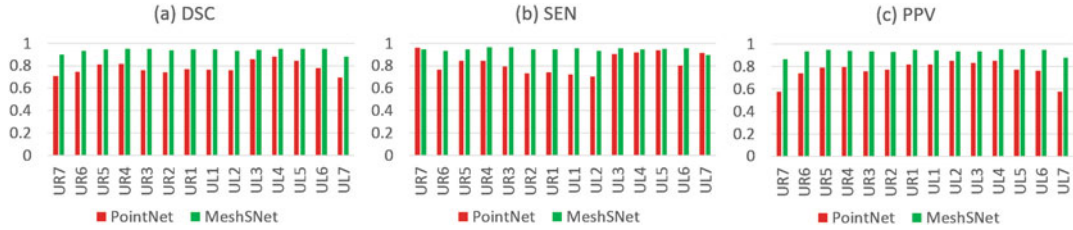


Fig. 10.5 Segmentation results for each of the 14 teeth (i.e., UR7 to UL7) quantified in terms of three evaluation metrics (i.e., DSC, SEN, and PPV), under threefold cross-validation. Reprinted from Ref. [18] with permission from Springer, Cham

between PointNet and MeshSegNet. All p-values in terms of the three metrics were small (i.e., 1.6×10^{-10}, 8.1×10^{-7}, and 6.1×10^{-9}), indicating that the segmentation performances of MeshSegNet were improved significantly. Moreover, compared with PointNet, MeshSegNet

effectively reduced false positives for segmenting molars and false negatives for segmenting central incisors (the first and second rows of Fig. 10.4). Notably, MeshSegNet annotated molars more precisely even though molars were not completely captured by IOS. (See areas indicated by green circles in the third row of Fig. 10.4.) The results of individual tooth segmentation presented in Fig. 10.5 are consistent with the overall segmentation results summarized in Table 10.2. In Fig. 10.5, we can see that MeshSegNet yielded better DSC values than PointNet on all teeth (i.e., from UR7 to UL7). These results further verify the effectiveness of MeshSegNet in the task of automated tooth segmentation on a 3D dental surface. On the other hand, it is worth noting that the improvement brought by MeshSegNet compared with PointNet is relatively more significant for the segmentation of molars (e.g., UR7 and UL7). For example, the MeshSegNet improved the DSC from 0.711 to 0.900 (p-value $<10^{-4}$), and improved the PPV from 0.575 to 0.867 (p-value $<10^{-6}$) for segmenting UR7. Note that segmenting molars is a very challenging task, because they locate at deep intraoral regions and might not be completely captured during scanning. These results further suggest the robustness of the MeshSegNet method. Both the visual evaluation in Fig. 10.4 and the quantitative evaluation in Table 10.2 as well as Fig. 10.5 suggest that MeshSegNet is potentially useful in practice for automated tooth labeling on a dental surface.

10.3.4 Post-Processing

Although MeshSegNet provides accurate tooth segmentation results (approximate 93% compared to the ground-truth), it is still not satisfactory for use in clinical practice. Additionally, the task of segmentation done thus far is still under the simplified dental mesh model. A conversion for the segmentation results between the simplified and original resolutions is needed. For those interested in a detailed description, the following section outlines two post-processes to achieve those goals. Other readers may wish to proceed to the conclusion.

10.3.4.1 Label Refinement

Xu et al. [3] adopted the multi-label graph cuts method presented by Boykov et al. [24, 25] to refine the tooth segmentation results from their deep learning system. In the multi-label graph cuts method, an energy to minimize is given as follows:

$$E(L) = \sum_{i \in \mathcal{F}} D\left(p_i, l_i\right) + \lambda \sum_{(i,j) \in \mathcal{N}} V\left(p_i, p_j, l_i, l_j\right), \tag{10.5}$$

where cell i is labeled by the prediction model with l_i under the probability of p_i; D is a penalty function, defined as $D(p_i, l_i) = -\log(p_i(l_i))$; V is an interaction potential; $\mathcal{N}$ is a set of all pairs of neighboring points; and λ is a nonnegative constant. An interaction potential V is expressed as follows:

$$V\left(p_i, p_j, l_i, l_j\right) = \begin{cases} 0, l_i = l_j \\ -\log\left(\frac{\theta_{ij}}{\pi}\right)\phi_{ij}, l_i \neq l_j, \theta_{ij} \text{ is concave} \\ -\beta_{ij}\log\left(\frac{\theta_{ij}}{\pi}\right)\phi_{ij}, l_i \neq l_j, \theta_{ij} \text{ is convex} \end{cases}, \tag{10.6}$$

where $\beta_{ij} = 1 + \left|\hat{n_i} \bullet \hat{n_j}\right|$ and $\phi_{ij} = |c_i - c_j|$; $\hat{n_i}$ and c_i are the normal vector and the barycenter of cell i, respectively; and θ_{ij} is the dihedral angle between cell i and cell j. A label refinement/optimization problem can be achieved by minimizing Eq. 10.5. In addition, Eq. 10.6 enforces the optimization to favor concave, which is suitable for dental models whose boundaries (e.g., teeth–gingiva and tooth–tooth) are usually concave.

Xu et al. [3] suggested that the nonnegative constant λ of 20 and 100 are suitable for boundaries of teeth–gingiva and inter-teeth, respectively. We used $\lambda = 70$ to optimize the segmentation results from MeshSegNet on the simplified meshes, as shown in Fig. 10.6. After the label refinement process was applied to the unrefined dental mesh shown in Fig. 10.6a, it is easy to see that the most incorrect predicted labels were fixed (Fig. 10.6b). Moreover, these segmentation results are quite similar to those of the Ground-Truth (Fig. 10.6c). The DSC,

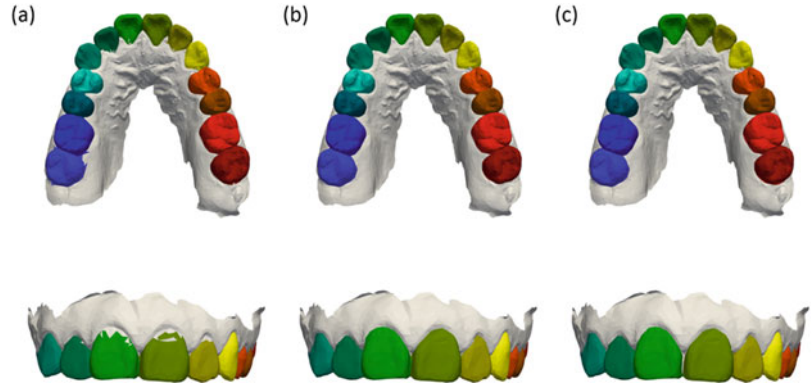


Fig. 10.6 (**a**) The segmentation results from MeshSegNet without label refinement. (**b**) The segmentation results from MeshSegNet with label refinement using multi-label graph-cuts algorithm. (**c**) The segmentation results of Ground-Truth

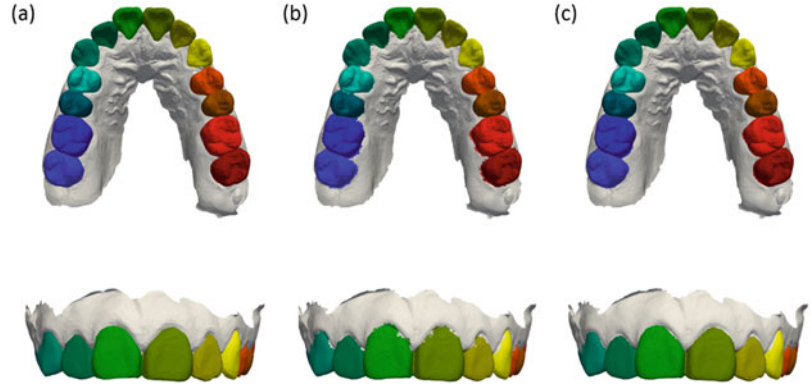


Fig. 10.7 The segmentation results on (**a**) a simplified mesh, (**b**) an original mesh using k-nearest neighbors (k-NN) with $k = 3$, and (**c**) an original mesh using support vector machine (SVM) with the radial basis function (RBF) kernel, respectively

SEN, and PPV were significantly improved to 0.987 ± 0.003, 0.992 ± 0.002, and 0.984 ± 0.005, respectively, showing that the multi-label graph cuts algorithm effectively optimizes the tooth segmentation results.

10.3.4.2 Label Mapping

The methods discussed above transfer accurate tooth segmentation results onto a simplified mesh. The final step is to project the segmentation result back onto the same mesh with original resolution. The goal of mapping labels on meshes with different resolutions is to train a model based on all cells and labels from a single simplified mesh. Then the model can predict labels on the original mesh. Unlike MeshSegNet using many meshes from different patients to train a model, an individual patient has its own model for the task of mapping. Since the two meshes are from the same patient, this idea is similar to "overfitting" but is perfect for the task of label mapping.

The k-nearest neighbors (k-NN) algorithm for classification is the simplest way to achieve the goal. The k-NN algorithm is a type of lazy learning method, which does not attempt to "learn" a model but simply stores the training data. The original k-NN classification rule performs prediction for a given sample by a simple majority vote of its k (e.g., $k = 3$) nearest neighbors in the training pool. An advanced candidate method for the mapping task is the support vector machine (SVM). An SVM learns an optimal separating (i.e., maximum margin) hyperplane so that the hyperplane can split different categories on the new samples. Both the k-NN and SVM have several versions (e.g., nonlinear kernel for SVM) but they are not sensitive to affect the result. The k-NN and SVM can be easily implemented by using a Python package *scikit-learn* (scikit-learn.org/stable/).

Figure 10.7a, b, and c show the results of segmentation on a simplified mesh, segmentation on an original mesh using k-NN with $k = 3$, and segmentation on an original mesh using SVM with the radial basis function (RBF) kernel, respectively. The processing time of k-NN and SVM is

0.90 s and 42.29 s, respectively, using the Intel® i7-7820X process. However, the SVM with RBF kernel exhibits a slightly better result, compared to the results from the k-NN.

10.4 Conclusion

CAD/CAM plays an important role in modern dentistry, particularly in orthodontics. In this chapter, we introduced a method for automated tooth segmentation on 3D dental models. Automated tooth segmentation is a difficult task due to the high variation in the teeth of different patients. Although several tooth segmentation methods exist, most of them are either labor-intensive or do not provide the necessary accuracy. Recently, machine (deep) learning has displayed an impressive performance in several different types of imaging such as 2D/3D images and graphs. We comprehensively reviewed a deep learning method, MeshSegNet, for implementing automated tooth segmentation, from preprocessing to post-processing. The segmentation results using MeshSegNet have a Dice similarity coefficient of 0.938 ±0.006 and even achieve 0.987 ±0.003 after refinement processing. Despite the fact that we only introduced the task of tooth segmentation here, the variant of MeshSegNet for automatically detecting landmarks (e.g., contact points) will also be developed and will play a crucial role in advanced CAD/CAM systems of the future.

References

1. Martin CB, Chalmers EV, McIntyre GT, et al. Orthodontic scanners: what's available? J Orthod. 2015;42:136–43. https://doi.org/10.1179/1465313315Y.0000000001.
2. Liao S, Liu S, Zou B, et al. Automatic tooth segmentation of dental mesh based on harmonic fields. Biomed Res Int. 2015;2015:187173.
3. Xu X, Liu C, Zheng Y. 3D tooth segmentation and labeling using deep convolutional neural networks. IEEE Trans Vis Comput Graph. 2019;25:2336–48. https://doi.org/10.1109/TVCG.2018.2839685.
4. Bronstein MM, Bruna J, LeCun Y, et al. Geometric deep learning: going beyond Euclidean data. IEEE Signal Process Mag. 2017;34:18–42. https://doi.org/10.1109/MSP.2017.2693418.
5. Maturana D, Scherer S. VoxNet: a 3D Convolutional Neural Network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 922–928.
6. Wu Z, Song S, Khosla A, et al. 3D ShapeNets: a deep representation for volumetric shapes. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1912–1920.
7. Qi CR, Su H, Nießner M, et al. Volumetric and multi-view CNNs for object classification on 3D Data. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5648–5656.
8. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs, 2013. arXiv Prepr arXiv13126203.
9. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in neural information processing systems, 2016, pp. 3844–3852.
10. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks, 2016. arXiv Prepr arXiv160902907.
11. Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks, 2019. arXiv Prepr arXiv190100596.
12. Zhou J, Cui G, Zhang Z, et al. Graph neural networks: a review of methods and applications, 2018. eprint arXiv:1812.08434 arXiv:1812.08434.
13. Monti F, Boscaini D, Masci J, et al. Geometric deep learning on graphs and manifolds using mixture model CNNs, 2016. eprint arXiv:1611.08402 arXiv:1611.08402.
14. Qi CR, Su H, Mo K, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation, 2016. eprint arXiv:1612.00593 arXiv:1612.00593.
15. Ko C-C, Tanikawa C, Wu T-H, et al. Machine learning in orthodontics: application review. Moyers Symp under revi, 2019.
16. Huang Q, Wang W, Neumann U. Recurrent slice networks for 3D segmentation of point clouds, 2018. eprint arXiv:1802.04402.
17. Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems, 2017. pp. 5099–5108.
18. Lian C, Wang L, Wu T-H, et al. MeshSNet: deep multi-scale mesh feature learning for end-to-end tooth labeling on 3D dental surfaces BT – Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. In: Shen D, Liu T, Peters TM, et al., editors. MICCAI 2019. Cham: Springer; 2019. p. 837–45.
19. Lian C, Wang L, Wu T, et al. Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3D intraoral scanners. IEEE Trans Med Imaging. 2020;39:2440–50. https://doi.org/10.1109/TMI.2020.2971730.

20. Schroeder WJ, Lorensen B, Martin K. The visualization toolkit: an object-oriented approach to 3D graphics. Kitware; 2004.
21. Gardner MW, Dorling SR. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ. 1998;32:2627–36. https://doi.org/10.1016/S1352-2310(97)00447-0.
22. Sudre CH, Li W, Vercauteren T, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations BT – deep learning in medical image analysis and multimodal learning for clinical decision support. In: Cardoso MJ, Arbel T, Carneiro G, et al., editors. . Cham: Springer; 2017. p. 240–8.
23. Reddi SJ, Kale S, Kumar S. On the convergence of adam and beyond. 2019. eprint arXiv:1904.09237 arXiv:1904.09237.
24. Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. In: Proceedings of the seventh IEEE international conference on computer vision, 1999, vol. 1, pp. 377–384.
25. Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans Pattern Anal Mach Intell. 2004;26:1124–37. https://doi.org/10.1109/TPAMI.2004.60.

Assessment of Outcomes by Using Machine Learning

11

Shankar Rengasamy Venugopalan, Mohammed H. Elnagar, Deepti S. Karhade, and Veerasathpurush Allareddy

11.1 Introduction

In recent years, there has been a tremendous increase in the volume of data owing to availability of data elements from multiple sources, transitioning from paper-based records to electronic health record systems, increasing virtual storage capacities, and availability of extremely fast processing computers [1]. Medical and surgical specialties have been at the forefront of using artificial intelligence analytic strategies to leverage the advantages of "big data" to personalize treatment interventions to the needs of individual patients and assess outcomes [2–4]. Consequently, there has been much improvement in the overall healthcare while reducing costs of delivery of care. In parallel with medical and surgical specialties, the field of Dentistry has made rapid strides in using artificial intelligence technologies to assess clinical outcomes. One of the artificial intelligence tools that is being increasingly used to analyze the electronic health record systems is machine learning. In the following sections, we will provide an overview of machine learning and how it can be applied in assessing outcomes. Specifically, we will focus on recent advances in craniofacial genomics and outcomes pertaining to images.

11.2 Overview of Machine Learning

Machine learning is a scientific discipline which seeks to emulate human intelligence by allowing computers to use past data in order to predict future data. As the workhorse in the era of big data, machine learning applications have revo-

S. R. Venugopalan
The University of Iowa College of Dentistry and Dental Clinics, Iowa City, IA, USA

M. H. Elnagar
University of Illinois at Chicago College of Dentistry, Chicago, IL, USA

D. S. Karhade
University of North Carolina at Chapel Hill Adams School of Dentistry, Chapel Hill, NC, USA

V. Allareddy (✉)
University of Illinois at Chicago College of Dentistry, Chicago, IL, USA

Head of Department of Orthodontics, University of Illinois at Chicago College of Dentistry, Chicago, IL, USA
e-mail: sath@uic.edu

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_11

lutionized the fields of medicine, engineering, finance, entertainment, and computational biology [5]. There are two broad categories of machine learning: supervised and unsupervised machine learning. Supervised learning builds a model by classifying each input into known classes, whereasunsupervised learning infers patterns and unravels the underlying structure of an unlabeled training dataset.

Supervised Machine Learning Supervised machine learning describes the process whereby a machine learning algorithm is trained to predict outcomes by using key defining features in the input variables. If the outcome is a continuous variable, the resulting model is known as a regression. If the supervised learning problem consists of discrete classes, the model is called a classification. Support Vector Machine (SVM) is a powerful model that structurally handles binary problems by solving a constrained quadratic optimization problem [6]. SVM works by mapping data within a high-dimensional space such that data points can be categorized even when they exist in a nonlinear fashion. By creating a hyperplane, a straight line within the high-dimensional space that separates two distinct classes, SVM can predict labels based on different feature vectors [7]. SVM can be used to identify fraudulent credit card activity, classify microarray gene expression profiles, and recognize handwritten numbers [8].

Logistic regression, decision trees (DT), and artificial neural networks (ANN) all yield both a class label and probability of being included in a class membership [9]. A logistic regression is performed when there are one or more independent variables that determine a dichotomous outcome. In the process of finding the line of best fit to describe how the set of independent variables relate to the dichotomous outcome variables, the logistic regression is able to generate coefficients and predict the logit transformation of the probability of presence and absence of characteristics of interest.

The decision tree methodology classifies a dataset into the structure of an inverted tree and is an effective way to handle large, complex datasets without a parametric structure [10]. Decision trees classify data based on posing a series of yes or no questions, such that each internal node points to a child node representing "yes" or "no." Thus, every question is represented as a hierarchy starting from the root, or the topmost node, to the leaf, which represents the last node in the DT, without children [11]. Because decision trees maximize the separation of data and produce more understandable rules compared to many other machine learning methods, they are less of a black box and are thus widely used in medicine [9, 12]. For instance, a DT can be used to diagnose which disease a patient has based on their presentation of symptoms [10]. The random forests method is an aggregation of numerous decision trees, where each tree is a standard Classification or Regression Tree (CART). The random forest method results in a reduction of variance when compared to a single decision tree [13]. This is an example of "ensemble learning." Each tree is built using CART and Decrease Gini Impurity to determine how to split the data9. Random forests are frequently used in the analysis of multi-omics data because of their ability to capture nonlinear associations between predictors and responses [14].

Artificial neural networks are computational models which are used for classification and consist of several highly interconnected neurons operating parallelly [9]. The structure of an ANN models the human brain. Artificial neural networks are organized into layers and consist of an input layer, several intermediate hidden layers and an output layer. Each layer consists of several interconnected nodes, analogous to neurons, which contain an activation function that transforms the output of each node within a layer. Backpropagation neural networks undergo a supervised learning process that allows forward activation of output flow and subsequent backpropagation of weight adjustments. When presented with a pattern, the ANN initially guesses what the pattern might be by randomly assigning weights. Later, as it learns how its prediction relates to the true outcome, it accordingly adjusts its connection weights to make more successful future predictions [9].

Unsupervised Machine Learning Unsupervised machine learning is used to identify naturally occurring patterns within a dataset [15]. Clustering, a technique frequently used in bioinformatics, partitions data into coherent groups. For example, K-means algorithms partition data such that K numbers clusters are present, and each data value lies closer to the mean of its cluster than any other. In contrast, hierarchical clustering is often used when separate clusters are not found within the dataset. In hierarchical clustering, multiple rounds of clustering are performed such that clusters are merged or divided at each round. Other forms of unsupervised learning include outlier detection, dimensionality reduction, and quantile estimation [16]. Semi-supervised machine learning is the intermediate step between supervised and unsupervised machine learning and is especially useful when labeled data is scarce or expensive to obtain [16]. Examples of such models include self-training, semi-supervised support vector machines, and mixture models [17].

11.3 Craniofacial Genomics Using Machine Learning

Craniofacial development is an intricate process under tight genetic and epigenetic control, which involves cell proliferation, differentiation, and interaction in a spatiotemporal fashion [18]. The gene expression process, governing embryonic development or homeostasis, begins with a complex signaling cascade triggering transcription of mRNA from the genomic blueprint (DNA). The transcribed mRNAs are spliced, processed, and translated to proteins, which executes the cellular function. The epigenetic modifications such as methylation of cytosine residues in the DNA or posttranslational modifications of histones add a layer of complexity to the gene expression process. In addition, the small RNA molecules such as micro RNAs and long noncoding RNAs fine-tune the gene expression and makes the gene regulatory network highly complex. In this highly coordinated process, any mishaps, depending on its magnitude, would result in a deviation from normal development, which could manifest phenotypically either as a simple variation or severe developmental defect. Before the advent of next generation sequencing, identifying gene(s) associated with congenital defects, required decades of dedicated and concerted effort of clinicians and scientists. However, decoding the human genome has catapulted the sequencing technology in a dramatic fashion, resulting in a rapid decline of sequencing cost from US$10 million in 2007 to $1500 in 2015 [19]. Such rapid decline in sequencing cost has greatly enhanced our abilities in identifying genes associated with population variation, congenital defects, and genetic diseases through genome wide association studies.

The high-throughput "omics" technologies allow us to study biological samples such as peripheral blood, whole saliva, biopsy tissues, or a single-cell isolated from human subjects and make "precision medicine" a possibility. The repertoire of high-throughput genomic (whole-genome or exome sequencing), transcriptomic (RNA Sequencing), epigenomic (Bisulfite Sequencing; Chromatin Immunoprecipitation Sequencing), and proteomic (Mass spectrometry) tools available today allows us to study complex gene regulatory network in a multidimensional view [20–24]. These technologies generate large volumes of data and often require sophisticated computational tools to analyze these large datasets. The recent advancements in the machine learning algorithms, a form of artificial intelligence, provides an opportunity to analyze the high-throughput "omics" data and derive meaningful information from these datasets [25].

In the field of genomics, machine learning strategies have a significant role to play in identifying genetic variants and establishing phenotype–genotype correlation [26]. Some of the major challenges in analyzing the whole-genome and -exome sequencing data are differentiating the disease-causing mutations from common genetic variations and predicting the effect of noncoding variants. The deep neural network machine learning algorithms have made a significant advancement in addressing these daunting tasks. In a recent report, Sundaram et al. have demonstrated that, by utilizing the common missense variants from nonhuman primate

species, the deep neural networks were trained to identify pathogenic mutations in patients with rare disease at 88% accuracy [27]. The deep learning based algorithm, DeepSEA, allows us to predict the effects of noncoding variants [28]. In addition, the deep learning algorithms allow to study the high-throughput functional assays and population genomic data in solving complex genomic sequence regions [29].

In executing normal cellular function, there are proteins that bind to DNA and/or RNA by identifying target sequence. The machine learning algorithms (DeepBind) have made a great progress in predicting the sequence specificities of proteins that bind to DNA and RNA [30]. During normal embryonic development, for proper cellular differentiation to occur, the enhancer region interacts with the promoter region. A novel natural language processing computational framework EP2vec allows us to predict enhancer-promoter interactions, which would greatly enhance our understanding of genetic regulation in health and disease [31]. In addition, these algorithms allow us to integrate the genomic, transcriptomic, epigenomic, and proteomic datasets to arrive at a comprehensive understanding of the biological mechanisms [32]. Epigenetic modification such as DNA methylation plays an important role in normal development and disease. The recent advancements allow us to identify DNA methylation changes in a single-cell resolution. The deep neural network machine learning algorithms such as DeepCpG enable us to analyze methylation data from a single-cell and make accurate predictions [33, 34].

11.4 Machine Learning Tools

There are several machine learning tools and libraries that have emerged in the last few decades for use in both academia and industry. To mention a few, Microsoft Excel is a tool with a variety of uses, not limited to numerical computation, data analysis, and visualization. IBM SPSS is menu-driven software for statistical analysis with built-in features for statistical computation, data mining, and building and visualizing predictive models. Both Microsoft and Excel are better suited for straightforward, smaller-scale data analysis which does not require background in programming and are very easy to use [35]. Stata, another software used for statistical analysis, is an intermediate step between user-friendliness of SPSS and requires less knowledge of pure coding when compared to R, SAS, and Python. MATLAB is a computing environment and programming language developed by MathWorks with included standard library and additional features available for purchase. Unlike MATLAB, both R and Python are open-source freely available statistical software commonly used to handle big data. Python is one of the easiest programming languages to begin learning and incudes several scientific libraries such as Pandas, to handle data and conduct time-series analysis and NumPy/SciPy for numerical and statistical computation [35]. Finally, R is a highly flexible statistical package with virtually unparalleled analytical power [35]. Because it is mostly utilized by businesses, its strength is analytics, rather than readability of output. For both R and Python, there are massive amounts of open-source coding available to users to get started when performing challenging data analysis.

11.5 Assessment of Images and Related Outcomes Using Machine Learning

Assessment of treatment outcomes requires analyses of various types of images like MRI (Magnetic Resonance Image), CT (Computed Tomography), cephalometric radiographs, panoramic radiographs, periapical radiographs, CBCT, 3D face scans, and 3D digital dental models. Medical and dental image analyses tend to be time consuming and require to be performed by a well-trained specialist (typically an Oral and Maxillofacial Radiologist). Multiple image processing steps are required to analyze, diagnose, and evaluate treatment outcomes. Some of the image processing steps include image orientation, digitization of anatomical landmarks,

segmentation, and extraction of data. Recently, there has been much interest in using machine learning techniques to evaluate medical and dental images [36, 37]. Machine learning is being increasingly used as a tool to help physicians to interpret medical imaging findings and reduce interpretation times [38].

11.5.1 2D Lateral Cephalometric Radiographs

Cephalometric radiographs are an essential tool for orthodontic diagnosis, treatment planning, and evaluation of end of treatment outcomes. Identification of anatomical landmarks on the lateral cephalometric radiographs is the first step for interpretation. Several cephalometric analyses have been developed based on various combinations of linear and angular measurements [39]. Manual cephalometric analyses are typically time consuming and need to be performed by a well-trained expert. Furthermore, significant interobserver variability has been observed with manual lateral cephalometric radiograph assessments [40]. During the past decade, several attempts have been made to develop an automated computerized system for the detection of anatomical landmarks and cephalometric analyses [41–43]. However, due to the complexity of the process, the developed methods were not of sufficient quality to replace manual digitizing [44]. With recent advances in machine learning techniques, more promising systems have been introduced. Lindner et al. introduced a fully automatic landmark annotation (FALA) system for identifying cephalometric landmarks and classifying skeletal malformations [44]. The FALA system consisted of a machine learning approach where Random Forest regression voting was used to detect the position, scale, and orientation of the skull and then used the Constrained Local Model framework (RFRV-CLM) to locate 19 landmarks in 24 seconds. The overall average point-to-point error was 2.2 ± 0.03 mm (Fig. 11.1).

Arik and colleagues proposed a fully automated framework for cephalometric analysis using deep learning techniques named convolutional neural networks [CNN] [45]. CNNs are biologically inspired variants of multilayer perceptron type of deep machine learning techniques. They are efficient for image processing and recognition applications because they exploit the spatial local correlation by imposing local connectivity patterns [45]. CNNs have demonstrated a wide range of image processing applications. However, the downside is that there is a need for abundant data for training [46].

In 2019, Kunz and colleagues created an automated cephalometric radiographic analysis using a customized CNN deep learning algorithm. The training data set consisted of 1792 different cephalometric images digitized by experienced examiners. To evaluate the quality of algorithm, each examiner analyzed 12 commonly used orthodontic parameters based on 50 cephalometric radiographs that were not part of the training data for the CNN (Fig. 11.2). The investigators were able to generate an algorithm capable of analyzing new cephalometric radiographs with comparable precision to experienced human examiners in a fraction of a second [47].

11.5.2 3D Cephalometric Radiographs

The standard 2D cephalometric radiograph has many limitations as it is a projection of the skull, which is a highly sophisticated 3D object onto a single 2D plane, leading to overlapping structures. Besides, head positioning variations during image acquisition and radiographic distortion cause the left and right outlines to not be perfectly superposed [48]. Three-dimensional cephalometric analyses with computed tomographic (CT) images and cone-beam computed tomography (CBCT) is gaining in popularity because it overcomes the limitations of traditional 2D radiographs. Furthermore, it is especially beneficial when there is facial asymmetry or a need for an orthognathic surgery. Developing an automated machine learning based 3D cephalometric analysis framework is challenging

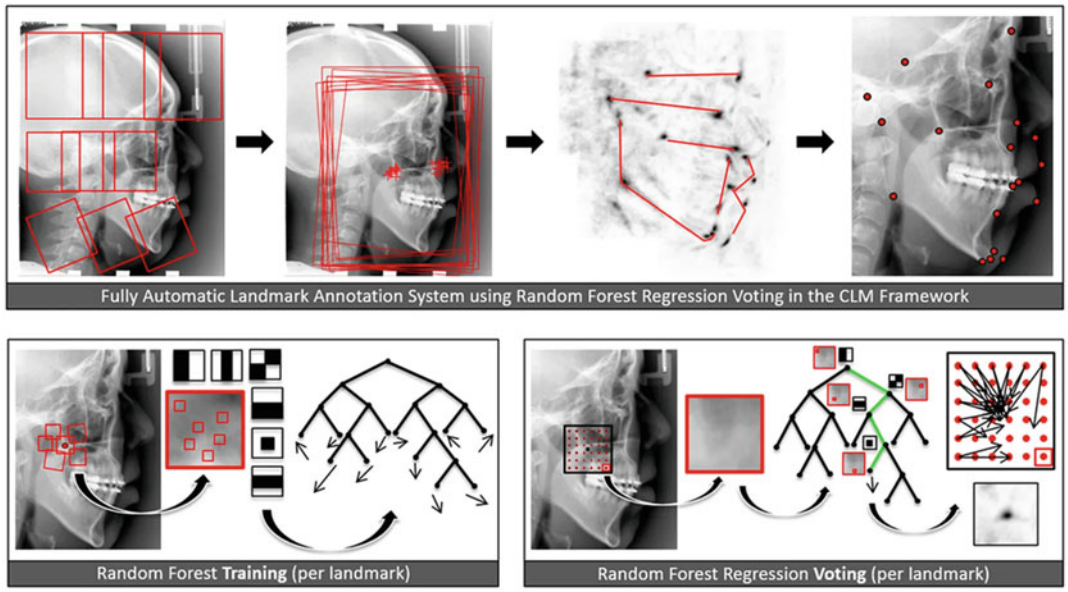


Fig. 11.1 Schematic overview of the FALA system [44]. Permission to use this figure was obtained from—Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully Automatic System for Accurate Localisation and Analysis of Cephalometric Landmarks in Lateral Cephalograms. Sci Rep. 2016 Sep 20;6:33581. doi: https://doi.org/10.1038/srep33581

because of multifold reasons including the 3D volume data has an increased number of parameters; there is a need for intensive computing resources; and greater computational complexity [49]. Lee and colleagues proposed an automatic 3D cephalometric annotation system which used a shadowed 2D image-based machine learning strategy [49]. It takes advantage of multiple 2D images containing 3D geometric morphology of the skull surface in x dimension, where x represents 3D volume data [49]. The representation of 3D images by 2D ones not only allows learning with relatively small data numbers but also solves the out-of-memory problem [49]. However, this group of investigators were able only to localize seven landmarks.

11.5.3 CBCT and CT Auto-Segmentation

Accurate CBCT and CT segmentation is a necessity for implant surgery, orthodontic/orthognathic surgery diagnosis and treatment planning, assessment of treatment outcomes, and dental forensics. CBCT segmentation includes segmentation of teeth, bones as well as soft tissue. The manual segmentation process is time consuming (Figs. 11.3, 11.4 and 11.5). The automated segmentation process would immensely benefit clinicians and researchers. However, the automated jaw bone segmentation on CBCT images using mathematical models is challenging for many reasons: the teeth usually have a similar intensity as the bone; large variations in the appearance of anatomical structures; and inconsistent manual labeling can confuse the trained models for segmentation [50, 51]. To solve this problem, methods that incorporate prior information about the expected shapes and position of the segmentation objects have been proposed [52].

Qiu et al. proposed a convolutional neural networks (CNNs) based approach for 3D mandible segmentation in CT scans [51]. Such an approach takes a multiplanar volume-to-slice strategy, which takes into account the spatial information of adjacent slices in order to preserve

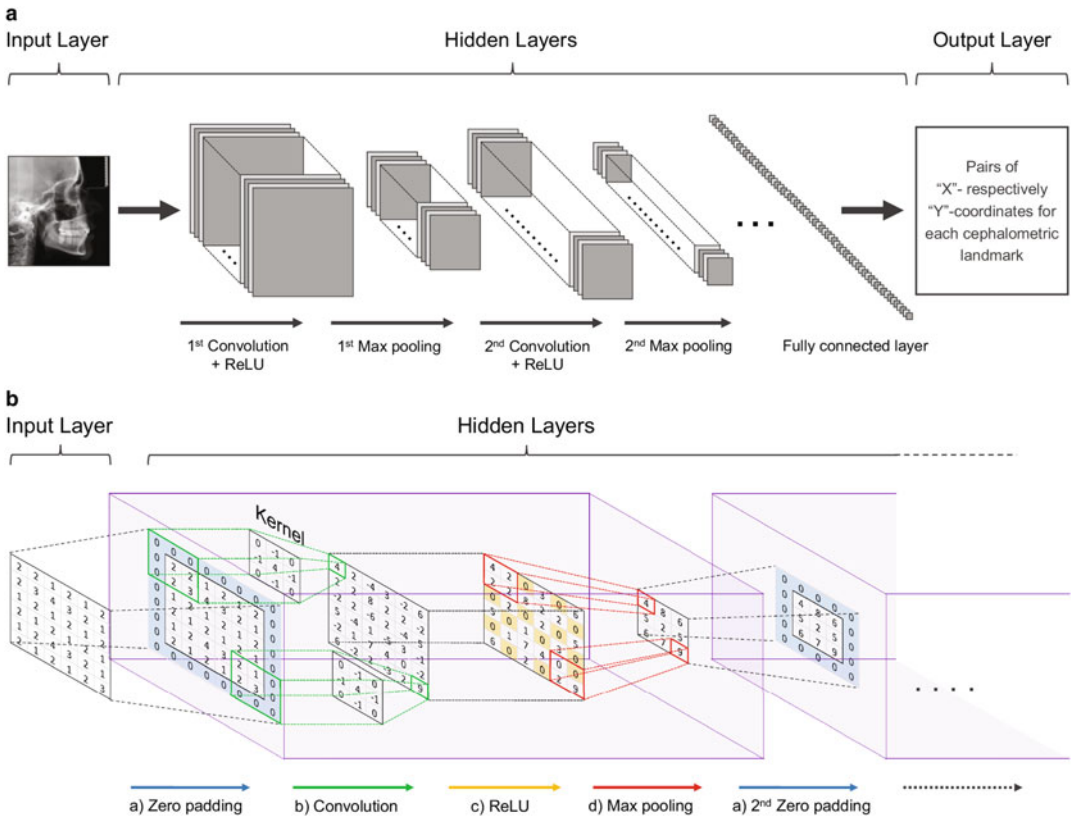


Fig. 11.2 (**a**) An Illustration of the convolutional neural network (CNN) design used to analyze cephalometric X-rays. (**b**) Schematic illustration of convolution and pooling processes in a CNN [47]. Permission to use this figure was obtained from—Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics: Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. J Orofac Orthop. 2020 Jan;81(1):52–68. doi: https://doi.org/10.1007/s00056-019-00203-8. Epub 2019 Dec 18

the connectivity of anatomical structures. They used 52 cases as training, 8 cases as validation, and 49 cases as a test. They could achieve an average Dice coefficient of 0.881 and an average surface error of 0.5791 mm on 109 CT scans and an average Dice coefficient of 0.9328 and 95HD of 1.4333 mm on the Public Domain Database for Computational Anatomy (PDDCA) dataset [51]. Their results demonstrated the effectiveness of the proposed approach in mandible segmentation and its potential use in 3D virtual planning of craniofacial tumor resection and free flap reconstructions.

CBCT-based image analysis plays a significant role in the detection and assessment of Craniomaxillofacial (CMF) deformities, which require precise segmentation of bones of the craniofacial region. Despite recent efforts for making a fully automated and accurate segmentation of bones and landmarking, the problem remains mostly unsolved especially for those with congenital or developmental deformities for whom precise diagnosis and treatment planning are most critically needed [51]. Torosdagli et al. reported a fully automated image analysis software based on Deep Geodesic Learning for mandible segmentation and anatomical landmarking that can overcome the highly variable clinical phenotype in the CMF region [53].

Acquisition and segmentation of complete 3D teeth models are essential for digital dentistry. For example, in orthodontics, it is needed for diagnosis and treatment planning, construction

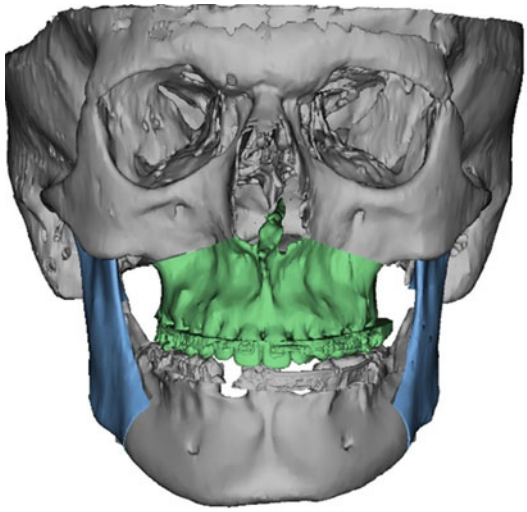


Fig. 11.3 CBCT segmentation for upper and lower jaw orthognathic surgery. Virtual treatment planning

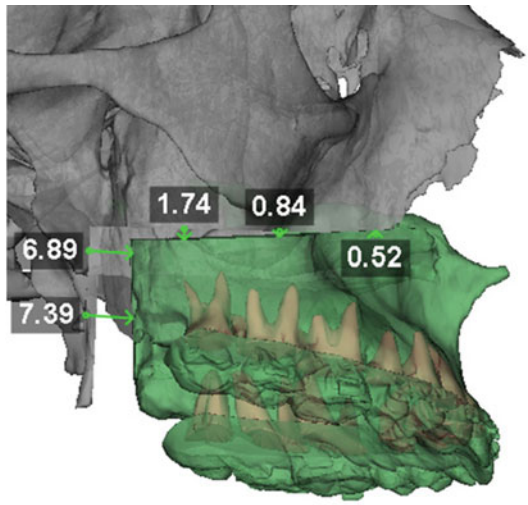


Fig. 11.4 Maxilla and maxillary teeth segmentation from CBCT

of customized orthodontic appliances, surgical guides for orthognathic cases, and assessment of treatment outcomes [54]. CBCTs can be used for acquiring complete 3D teeth models. However, segmenting teeth from CBCT images is a difficult problem for the following reasons: it is hard to separate lower teeth from its opposing upper teeth along their occlusal surface because of the lack of changes in gray values if CBCT is acquired in occlusion; it is hard to separate a tooth from its surrounding alveolar bone because of their highly similar densities; and adjacent teeth have a similar shape. Cui and colleagues proposed a new machine learning based method for automatic tooth segmentation and identification [55]. Their method consisted of a two-stage deep supervised neural network. In the first stage, to enhance the boundary of blurring and low contrast signals, they trained an edge extraction subnetwork. In the second stage, they devised a 3D region proposal network with a novel learned similarity matrix which efficiently removes the duplicated proposals, speeds up and stabilizes the training process, and significantly cuts down the GPU memory usage [55].

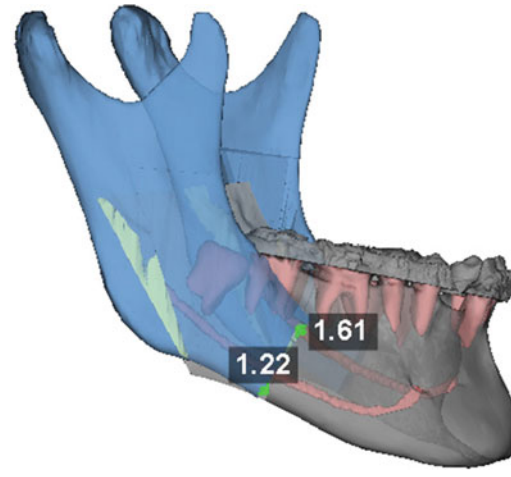


Fig. 11.5 Mandible and mandibular teeth segmentation from CBCT

11.5.4 Digital 3D Dental Models

Digital 3D dental models have replaced traditional plaster models in dental offices. The digital 3D dental models facilitate the measurement of a tooth position in the three dimensions, can be manipulated, sectioned, and superimposed to assess the treatment outcome [56]. Furthermore, digital 3D dental models are fundamental for computer-aided design (CAD) dental systems, fabricating customized orthodontic appliances, and virtual surgical planning. Tooth segmentation and labeling is the most critical component of these CAD

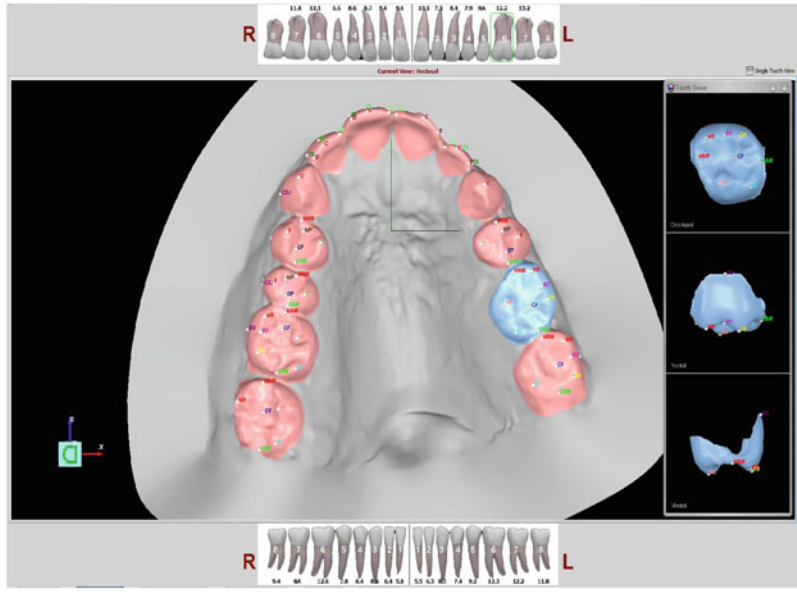


Fig. 11.6 Teeth manual segmentation on digital 3D dental models

systems and digital systems (Fig. 11.6). Manual segmentation and manipulation of the 3D digital models is time-consuming and needs a trained expert [54]. Having an automated process will reduce the time and effort needed to create CAD dental restorations, implants, customized digital orthodontic appliances, and virtual surgical planning, besides facilitating assessment of treatment outcomes.

Xu et al. proposed a data-driven method for 3D dental mesh segmentation by using deep CNN model [57]. The network was designed to provide automated labeling for each tooth triangle. It receives a detailed 3D dental model as input and outputs the label list of each triangle face. They suggested a label-free mesh simplification method for preserving teeth boundary information and significantly improved the efficiency of the networks. They claimed that their model achieved 99.06% accuracy for the maxillary arch model and 98.79% for the mandibular arch model [57].

11.5.5 Remote Dental Monitoring

With advances in telecommunication and information technology, the field of dentistry begun its foray into "teledentistry" to overcome barriers to access to care [58]. Furthermore, the advent of smartphones made two-way real-time patient–doctor interaction easier [59]. Interest in remote treatment monitoring has also led to the creation of orthodontic smartphone applications (apps) specially designed for this purpose. One of the more advanced apps currently available called Dental Monitoring™ (DM™) is a digital technology software that allows orthodontists to monitor patients remotely through continuous analytics using control vision technology, metaheuristics, and artificial Intelligence. This app utilizes a patented artificial intelligence machine learning algorithm to calculate 3-dimensional (3D) tooth movements from intraoral photos and videos that patients capture using their smartphone cameras. The accuracy error claimed by DM™ is less than 0.1 mm and less than 0.5 degrees for tip and torque [60].

The DM™ app comprises three interconnected platforms: a smartphone application for patients; a patented tooth movement artificial intelligence tracking algorithm; and an online Doctor Dashboard® where orthodontists can view patient treatment progress and pre- or post-treatment changes (Fig. 11.7). To begin using DM™

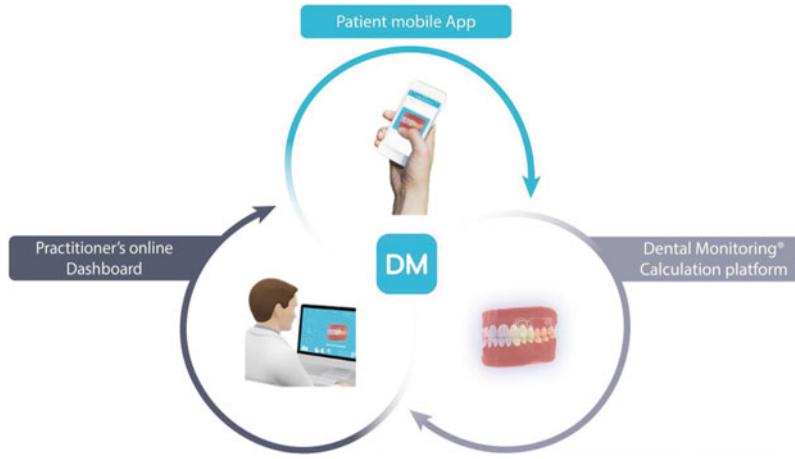


Fig. 11.7 Dental Monitoring consists of three integrated platforms: a mobile application for the patient, a patented movement tracking algorithm, and a web-based Doctor Dashboard

Fig. 11.8 Dental Monitoring scan box attached to smartphone

technology, the orthodontist uploads an initial 3D digital model in the STL file format to the Doctor Dashboard® [54]. In addition, the patient takes a video or does a photo examination through the DM™ application on their smartphone using a DM™ patented cheek retractor and scan box (Fig. 11.8). The DM™ app uses the initial scan and pretreatment video or photo examination to establish a baseline of tooth position and occlusion from which to calculate future movements. The results are then viewable on the Doctor Dashboard® in the form of graphs, photos, and a unique 3D visualization of the current tooth position called 3D Matching™. The 3D Matching™ is a feature of DM™ app that creates a multidimensional information map of the teeth to allow for replay of tooth movement (Fig. 11.9) [60, 61]. In addition to tracking tooth position, DM™ app also tracks patients' oral hygiene and sends automated alarms to the patients and doctor. Furthermore, the DM™ app claims that their artificial intelligence algorithm can auto-detect clinical situations such as broken brackets, wires, and unseated

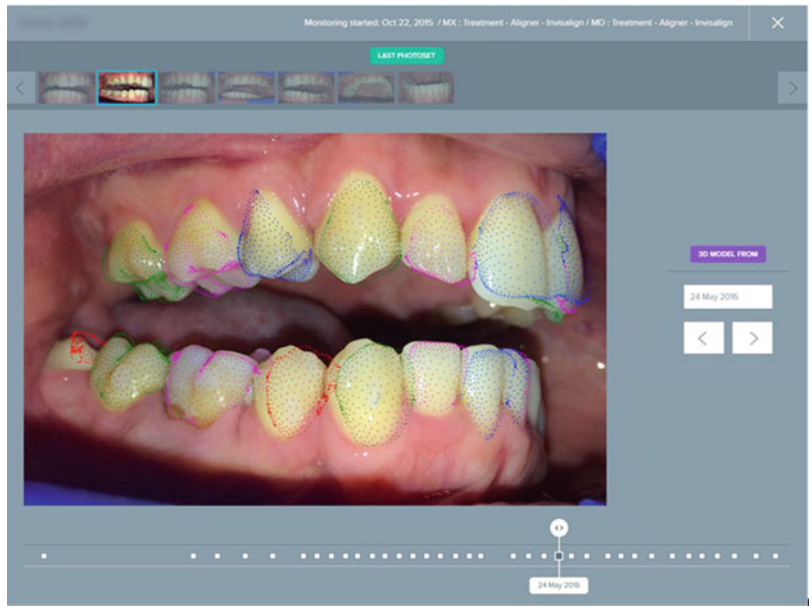


Fig. 11.9 3D matching for the evolution of the treatment effects and tracking teeth movement with clear aligner

clear aligners [60]. DM™ technology could help orthodontists reach their treatment goals more frequently and efficiently by alerting the practitioner of changing conditions as they occur or when pretreatment objectives are close to being reached, thus optimizing the timing of patient appointments and preventing unnecessary ones. But more studies need to be done to validate this technology.

11.6 Conclusions

In essence, the machine learning holds a great promise for "personalized medicine and dentistry." The artificial intelligence algorithms are well poised to enhance our understanding of complex gene regulatory network in health and disease, which would be unimaginable a decade ago. This rapidly evolving technology has already shown immense potential in analyzing the multi-omics data, which is a great step forward in understanding the genetic make-up of our patients in personalizing treatment and assessing outcomes. In addition, the artificial intelligence algorithms demonstrate promise in analyzing and interpreting medical/dental radiographs and images in facilitating faster diagnosis and providing personalized treatment. This technology is for here to stay, evolve rapidly, and perhaps revolutionize the way we practice medicine and dentistry.

References

1. Allareddy V, Rengasamy Venugopalan S, Nalliah RP, Caplin JL, Lee MK, Allareddy V. Orthodontics in the era of big data analytics. Orthod Craniofac Res. 2019 May;22(Suppl 1):8–13. https://doi.org/10.1111/ocr.12279.
2. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 2018 Jun 12;71(23):2668–79.
3. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014 Feb 7;2:3.
4. Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. Front Genet. 2015 Jun 30;6:229.
5. El Naqa I, Li RJ, Murph MJ, editors. What is machine learning? Cham: Springer; 2015.
6. Bastanlar Y, Ozuysal M. Introduction to machine learning. Methods Mol Biol. 2014;1107:105–28.
7. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in Cancer genomics. Cancer Genomics Proteomics. 2018;15:41–51.

8. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24:1565–7.
9. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002;35:352–9.
10. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27:130–5.
11. Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008;26:1011.
12. Cortez P, Embrechts MJ. Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci. 2013;225:1–17.
13. Couronné R, Probst P, Boulesteix A-L. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinform. 2018;19:270.
14. Manor O, Zubair N, Conomos MP, Xu X, Rohwer JE, Krafft CE, et al. A multi-omic association study of Trimethylamine N-oxide. Cell Rep. 2018;24:935–46.
15. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920–30.
16. Chapelle O, Schlkopf B, Zien A. Semi-supervised learning. Cambridge, MA: The MIT Press; 2010.
17. Zhu X, Goldberg AB. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning. 2009; 3:1–130.
18. Twigg SR, Wilkie AO. New insights into craniofacial malformations. Hum Mol Genet 2015 Oct 15;24(R1):R50–R59. https://doi.org/10.1093/hmg/ddv228. Epub 2015 Jun 17.
19. Payne K, Gavan SP, Wright SJ, Thompson AJ. Cost-effectiveness analyses of genetic and genomic diagnostic tests. Nat Rev Genet 2018 Apr;19(4):235–246. https://doi.org/10.1038/nrg.2017.108. Epub 2018 Jan 22.
20. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: history and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J 2019 Nov 17;18:9–19. https://doi.org/10.1016/j.csbj.2019.11.002. eCollection 2020.
21. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017 Aug 18;9(1):75. https://doi.org/10.1186/s13073-017-0467-4.
22. Jelin AC, Vora N. Whole exome sequencing: applications in prenatal genetics. Obstet Gynecol Clin N Am. 2018 Mar;45(1):69–81. https://doi.org/10.1016/j.ogc.2017.10.003.
23. Lundberg E, Borner GHH. Spatial proteomics: a powerful discovery tool for cell biology. Nat Rev Mol Cell Biol. 2019 May;20(5):285–302. https://doi.org/10.1038/s41580-018-0094-y.
24. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. Nat Rev Genet 2015 Dec;16(12):716–726. https://doi.org/10.1038/nrg3980. Epub 2015 Oct 13.
25. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019 Jan;51(1):12–18. https://doi.org/10.1038/s41588-018-0295-5. Epub 2018 Nov 26.
26. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019 Nov 19;11(1):70. https://doi.org/10.1186/s13073-019-0689-8.
27. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, Fritzilas N, Hakenberg J, Dutta A, Shon J, Xu J, Batzoglou S, Li X, Farh KK. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 2018 Aug;50(8):1161–1170. https://doi.org/10.1038/s41588-018-0167-z. Epub 2018 Jul 23.
28. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015 Oct;12(10):931–934. https://doi.org/10.1038/nmeth.3547. Epub 2015 Aug 24.
29. Telenti A, Lippert C, Chang PC, DePristo M. Deep learning of genomic variation and regulatory network data. Hum Mol Genet. 2018 May 1;27(R1):R63–71. https://doi.org/10.1093/hmg/ddy115.
30. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015 Aug;33(8):831–838. https://doi.org/10.1038/nbt.3300. Epub 2015 Jul 27.
31. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics. 2018 May 9;19(Suppl 2):84. https://doi.org/10.1186/s12864-018-4459-6.
32. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell 2018 Jun 14;173(7):1581–1592. https://doi.org/10.1016/j.cell.2018.05.015. Epub 2018 Jun 7.
33. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017a Apr 11;18(1):67. https://doi.org/10.1186/s13059-017-1189-z.
34. Angermueller C, Lee HJ, Reik W, Stegle O. Erratum to: DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017b May 12;18(1):90. https://doi.org/10.1186/s13059-017-1233-z.
35. Ozgur C, Colliau T, Rogers G, Hughes Z, Myer-Tyson B. MatLab vs. Python vs. R. J Data Sci. 2017;15:355–72.
36. Currie GM. Intelligent imaging: anatomy of machine learning and deep learning. J Nucl Med Technol 2019 Dec;47(4):273–281. https://doi.org/10.2967/jnmt.119.232470. Epub 2019 Aug 10.
37. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, Kim N. Deep learning in medical imaging: general overview. Korean J Radiol 2017 Jul-Aug;18(4):570–584. https://doi.org/10.3348/kjr.2017.18.4.570. Epub 2017 May 19.

38. Summers RM. Improving the accuracy of CTC interpretation: computer-aided detection. Gastrointest Endosc Clin N Am. 2010 Apr;20(2):245–57. https://doi.org/10.1016/j.giec.2010.02.004.
39. Ferreira JTL, de Souza Telles C. Evaluation of the reliability of computerized profile cephalometric analysis. Braz Dent J. 2002;13(3):201–4.
40. Durão AP, Morosolli A, Pittayapat P, Bolstad N, Ferreira AP, Jacobs R. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study. Imaging Sci Dent 2015 Dec;45(4):213–220. https://doi.org/10.5624/isd.2015.45.4.213. Epub 2015 Dec 17.
41. Cardillo J. An image processing system for locating craniofacial landmarks. IEEE trans med imaging. IEEE Trans Med Imaging. 1994;13(2):275–89.
42. Grau V, Alcañiz M, Juan MC, Monserrat C, Knoll C. Automatic localization of cephalometric landmarks. J Biomed Inform. 2001 Jun;34(3):146–56.
43. Rudolph DJ, Sinclair PM, Coggins JM. Automatic computerized radiographic identification of cephalometric landmarks. Am J Orthod Dentofac Orthop. 1998 Feb;113(2):173–9.
44. Lindner C, Wang CW, Huang CT, Li CH, Chang SW, Cootes TF. Fully automatic system for accurate localisation and analysis of Cephalometric landmarks in lateral Cephalograms. Sci Rep. 2016 Sep 20;6:33581. https://doi.org/10.1038/srep33581.
45. Arik SÖ, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham) 2017 Jan;4(1):014501. https://doi.org/10.1117/1.JMI.4.1.014501. Epub 2017 Jan 6.
46. Yang X, Wu N, Cheng G, Zhou Z, Yu DS, Beitler JJ, Curran WJ, Liu T. Automated segmentation of the parotid gland based on atlas registration and machine learning: a longitudinal MRI study in head-and-neck radiation therapy. Int J Radiat Oncol Biol Phys 2014 Dec 1;90(5):1225–1233. https://doi.org/10.1016/j.ijrobp.2014.08.350. Epub 2014 Oct 13.
47. Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics: evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. J Orofac Orthop 2020 Jan;81(1):52–68. https://doi.org/10.1007/s00056-019-00203-8. Epub 2019 Dec 18.
48. Baumrind S, Frantz RC. The reliability of head film measurements. 1. Landmark identification. Am J Orthod. 1971 Aug;60(2):111–27.
49. Lee SM, Kim HP, Jeon K, Lee SH, Seo JK. Automatic 3D cephalometric annotation system using shadowed 2D image-based machine learning. Phys Med Biol. 2019 Feb 20;64(5):055002. https://doi.org/10.1088/1361-6560/ab00c9.
50. Fan Y, Beare R, Matthews H, Schneider P, Kilpatrick N, Clement J, Claes P, Penington A, Adamson C. Marker-based watershed transform method for fully automatic mandibular segmentation from CBCT images. Dentomaxillofac Radiol 2019 Feb;48(2):20180261. https://doi.org/10.1259/dmfr.20180261. Epub 2018 Nov 9.
51. Qiu B, Guo J, Kraeima J, Glas HH, Borra RJH, Witjes MJH, van Ooijen PMA. Automatic segmentation of the mandible from computed tomography scans for 3D virtual surgical planning using the convolutional neural network. Phys Med Biol. 2019 Sep 5;64(17):175020. https://doi.org/10.1088/1361-6560/ab2c95.
52. Indraswari R, Arifin AZ, Suciati N, Astuti ER, Kurita T. Automatic segmentation of mandibular cortical bone on cone-beam CT images based on histogram thresholding and polynomial fitting. Int J Intell Eng Syst. 2019;12(4):130–41. https://doi.org/10.22266/ijies2019.0831.13.
53. Torosdagli N, Liberton DK, Verma P, Sincan M, Lee JS, Bagci U. Deep geodesic learning for segmentation and anatomical Landmarking. IEEE Trans Med Imaging 2019 Apr;38(4):919–931. https://doi.org/10.1109/TMI.2018.2875814. Epub 2018 Oct 12.
54. Elnagar MH, Aronovich S, Kusnoto B. Digital workflow for combined orthodontics and Orthognathic surgery. Oral Maxillofac Surg Clin North Am 2020 Feb;32(1):1–14. https://doi.org/10.1016/j.coms.2019.08.004. Epub 2019 Nov 4.
55. Cui Z, Li C, Wang W. ToothNet: Automatic Tooth Instance Segmentation and Identification From Cone Beam CT Images, 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA, 2019, pp. 6361–6370. https://doi.org/10.1109/CVPR.2019.00653.
56. Elnagar MH, Elshourbagy E, Ghobashy S, Khedr M, Evans CA. Dentoalveolar and arch dimension changes in patients treated with miniplate-anchored maxillary protraction. Am J Orthod Dentofac Orthop. 2017 Jun;151(6):1092–106. https://doi.org/10.1016/j.ajodo.2016.10.038.
57. Xu X, Liu C, Zheng Y. 3D tooth segmentation and labeling using deep convolutional neural networks. IEEE Trans Vis Comput Graph 2019 Jul;25(7):2336–2348. https://doi.org/10.1109/TVCG.2018.2839685. Epub 2018 May 22.
58. Chen JW, Hobdell MH, Dunn K, Johnson KA, Zhang J. Teledentistry and its use in dental education. J Am Dent Assoc. 2003 Mar;134(3):342–6.
59. Aziz SR, Ziccardi VB. Telemedicine using smartphones for Oral and maxillofacial surgery consultation, communication, and treatment planning. J Oral Maxillofac Surg. 2009 Nov;67(11):2505–9. https://doi.org/10.1016/j.joms.2009.03.015.
60. Kravitz ND, Burris B, Butler D, Dabney CW. Teledentistry, do-it-yourself orthodontics, and remote treatment monitoring. J Clin Orthod. 2016 Dec;50(12):718–26.
61. Morris RS, Hoye LN, Elnagar MH, Atsawasuwan P, Galang-Boquiren MT, Caplin J, Viana GC, Obrez A, Kusnoto B. Accuracy of dental monitoring 3D digital dental models using photograph and video mode. Am J Orthod Dentofac Orthop. 2019 Sep;156(3):420–8. https://doi.org/10.1016/j.ajodo.2019.02.014.

Part IV

Machine Learning Supporting Dental Research

Machine Learning in Evidence Synthesis Research

12

Alonso Carrasco-Labra, Olivia Urquhart, and Heiko Spallek

In the following, we will explore how Machine Learning (ML) can support the labor intense process of conducting Systemic Reviews (SR). Given the importance of SRs in the process of developing evidence-based guidelines and there their potential impact on the health and well-being of large portions of the population, this is a topic of the utmost importance to researchers, clinicians, and patients. While we will provide a strong rationale in favor of using ML to increase the effectiveness and efficiency of developing SRs, we are also elaborating on the shortcomings and potential pitfalls of employing artificial intelligence (AI) to a process that has been mostly driven by human experts in the past. However, it is important to recognize that clinical decision-making does not follow a dichotomy of either human- or machine-based but sits at a continuum between the two extremes (Fig. 12.1). Educating the public and the health care professional community about these considerations is equally important to advancing the field of AI and using techniques such as ML for improving the process of developing SRs.

A. Carrasco-Labra (✉)
Department of Evidence Synthesis and Translation Research, American Dental Association Science and Research Institute (ADSRI), LLC., Chicago, IL, USA

Department of Oral and Craniofacial Health Science, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: carrascolabraa@ada.org

O. Urquhart
Department of Evidence Synthesis and Translation Research, American Dental Association Science and Research Institute (ADSRI), LLC., Chicago, IL, USA

H. Spallek
Head of the School and Dean, The University of Sydney School of Dentistry, Westmead, NSW, Australia

Academic Lead Digital Health and Health Service Informatics, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia

12.1 What Are Systematic Reviews?

Evidence-based clinical practice is rooted in the principle of using the best available evidence coupled with clinician expertise and patient values to inform clinical decisions [1, 2]. This evidence may be obtained from many sources (e.g., randomized controlled trials, and observational studies) and synthesized in a systematic or unsystematic manner. The method and type of information collected will impact how much trust

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_12

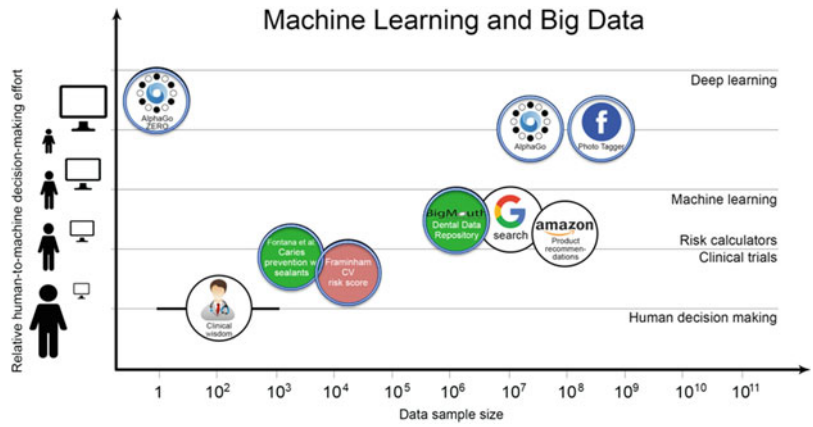


Fig. 12.1 Human-to-machine effort and relation to sample size. Adapted from: Beam AL, Kohane IS: Big Data and Machine Learning in Health Care. JAMA. 2018 Apr 3;319(13):1317–1318

we can place in the conclusions from the data. Systematic reviews (SRs) are one way to collect and synthesize the available evidence. They are regarded by the research community as one of the most reliable methods if conducted appropriately and following established methods. In the context of health care, a SR is a type of secondary research, in which all literature pertaining to a clinical question is collected according to pre-defined eligibility or selection criteria and synthesized for use in clinical practice ([3], p. 1). The methods are designed to ensure the reproducibility of the research and reduce bias.

Briefly, the SR process starts with the development of a research question. Next, an informationist (i.e., an individual with "the knowledge and expertise of a health care professional with the information retrieval skills of a librarian." [4]) searches multiple databases in an effort to capture all relevant literature. Then, the review authors read through the titles and abstracts of each article retrieved from the search and determine if it should be included or excluded from the SR. They then read full text articles of all included references and make a final decision as to whether or not to include in the SR. Subsequently, they extract relevant quantitative and qualitative data from each article, appraise it to determine the possibility for the study to be affected by bias or systematic error (i.e., risk of bias assessment). Finally, they synthesize data and provide estimates of the effect for a number of patient important outcomes, oftentimes using meta-analysis (Fig. 12.2).

Over time, SR methods have been adapted for a variety of reasons, including but not limited to:

1. Resource or time constraints necessitate a quick deliverable—often the case in public health emergencies [5].
2. Limited consensus exists about a topic, and the purpose of the review is to gain a broad understanding of the state of the literature [6].
3. A research question includes a complex social intervention [7].

In general, the workflow for these adapted methodological approaches (e.g., scoping reviews, rapid reviews) is similar to that of the "classic" SR. Completing each task is exceptionally labor intensive. Due to the sheer volume of literature that must be sifted through to find relevant studies, one project typically takes over a year to produce [8]. During this time, information becomes quickly outdated as the published literature base proliferates quickly. In 2010, for example, approximately 75 trials in the area of medicine were published per day. [9] We conducted a search in PubMed's Clinical Queries using the MeSH term "Dentistry," and revealed that conservatively, at least two trials were published each day in 2018 in the dental field. These inefficiencies are compounded by the fact that many tasks are completed manually and are carried out independently by two individualsin

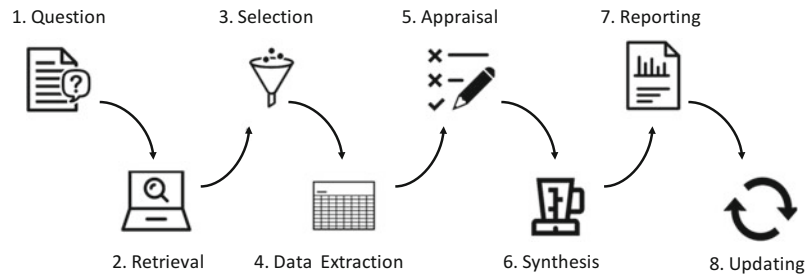


Fig. 12.2 Depiction of the steps in the systematic review process

an attempt to avoid human error. This raises the question: What if there is a way to maintain the rigorous standards associated with SRs while also increasing efficiency and reducing workload?

Automating the SR process was first introduced in the early 2000s [10, 11]. Not only does automation have the potential to increase efficiency in the process, but it can free up time to focus on more cognitively challenging and creative tasks. Integrating machine learning (ML) in SR development is one way to semi-automate the process, and to date, numerous tools have already been developed for this purpose [12]. It is important to note that, at the moment, not all tasks in the workflow are suitable for machine learning. Herein, we will summarize SR tasks, for which machine learning shows promising potential.

12.2 Research Question

Researchers conducting SRs start by defining an explicit and focused research question. Usually, they are required to identify the population or type of patients to include in the review (e.g., adults or children with a particular condition or characteristic), the experimental intervention of interest (e.g., surgical, pharmacological, and behavioral), a comparison intervention (e.g., another active intervention, placebo, no treatment, and standard of care), and the outcomes of interest (e.g., patient important outcomes or outcomes relevant to consumers) following the PICO mnemotechnic [13]. This research question guides the next steps in the review, which are the creation of the search strategy and evidence retrieval (Table 12.1).

12.3 Search

After the PICO question is clearly defined, the review authors will work with an informationist or someone trained in systematic searching to develop a search strategy. First, they identify all of the databases (e.g., PubMed, Embase, CINAHL) with literature relevant to their PICO or research question. Second, the informationist develops a search strategy to query the selected databases. Third, they test the strategy in one of the selected databases. The goal is to find all relevant articles (i.e., high recall or sensitivity), which may come at the expense of capturing many irrelevant articles (i.e., low precision or specificity) [14, 15]. Fourth, after a few rounds of refinement, the informationist reformats the search strategy to match the specifications (e.g., syntax) of each database. Finally, they combine the references retrieved from each database, and remove duplicates.

One of the applications of ML in searching is in study design filtering; more specifically identifying randomized controlled trials (RCT). When the purpose of a SR is to assess the effectiveness of an intervention or treatment (e.g., sealants and composite restorations), RCTs are regarded as the most appropriate study design. Furthermore, in the health care realm, SRs of interventions are pervasive, making study design filtering with the assistance of ML appealing. Existing tools use ML to predict the likelihood that an article is an RCT. RCT Tagger is a freely available tool in which a ML application allows reviewers to filter references from the database MEDLINE. After scanning the citation, abstract, and MeSHterms

Table 12.1 Examples of research questions in systematic reviews following the PICO format

		Components of the question			
Type of question	Research question	P = Patient or population	I = Intervention (Exposure or index test)	C = Comparison (or reference standard)	O = Outcomes
Diagnosis	How useful is an autofluorescence handheld device in detecting potentially malignant lesions?	Patients with a potentially malignant lesion in the mucosa	Use of autofluorescence handheld device	Biopsy and histopathological diagnostic	Diagnostic accuracy (true positive, true negative, false positives, and false-negative findings)
Etiology or harm	Is the exposure to a high volume of dental X-rays compared to low volume of dental X-rays associated with a higher risk of meningiomas?	Patients receiving dental X-rays	High volume of dental X-rays	Low volume of dental X-rays	Meningioma (brain tumor)
Therapy or prevention	What is the effect of 38% silver diamine fluoride solution compared to 5% NaF varnish in arresting or reversing noncavitated carious lesions?	Patients with noncavitated occlusal carious lesions	38% silver diamine fluoride solution	5% NaF varnish	Arrested or reversed carious lesion, toxicity
Prognosis	Are heavy-smoker patients receiving dental implants at higher risk of postoperative complications compared to nonsmoker patients?	Patients receiving dental implants	Heavy smokers	Nonsmokers	Dental implant failure, surgical site infection, bone resorption

associated with a reference, RCT Tagger calculates a probability for the article to be an RCT [16].

Similarly, RobotSearch is an open-access tool that, in addition to discerning RCTs from other study designs, enables users to choose if they would like the system to produce an output that maximizes "sensitivity/recall" or "specificity/precision" [17]. This function has great potential for those conducting rapid reviews, in which time is very limited, and a more specific, as supposed to a more sensitive search is valued. Lastly, Cochrane has an RCT filtering ML tool, available to their subscribers [18]. This tool combines crowdsourcing [19] and ML to identify RCTs. These tools may eventually replace existing RCT filters [20]. As of yet, there are no tools that filter by other study designs, such as non-randomized or diagnostic test accuracy studies [21].

ML has also been utilized to conduct "semantic searching." Typically, information specialists build search strategies with terms separated by Boolean operators (e.g., "AND," "OR"). These operators instruct the search engine to look for records with and without these terms. In contrast, semantic search engines identify concepts within a phrase, instead of reading a phrase as individual words strung together [22]. This is similar to searching by MeSH terms in PubMed [23]. MeSH terms are a collection of controlled vocabulary that were historically assigned manually to each article indexed in PubMed. Thalia is a semantic search engine that is capable of using ML to automatically assign MeSH terms to articles [24]. Most recently, PubMed has adopted an indexing approach in which a combination of ML and a set of rules is used to identify MeSH terms [25–27].

12.4 Study Selection/Screening

After the search is complete and references are de-duplicated, reviewers "screen" the title and abstract of each citation for potential eligibility in the SR. At this stage, they determine if it is worthwhile to continue reading the full text of an article with high chance of relevance. Eligibility judgments are based on pre-defined set of conditions, commonly referred to as selection or eligibility criteria. These criteria specify the study design, populations and patients' characteristics, interventions, or exposures that will be considered in the SR. Review authors are usually overinclusive at this stage, as sometimes the brief information in titles and abstracts does not enable a definitive conclusion about eligibility. After references highly suggestive of relevance are identified, authors retrieve their full text, most of the time in PDF format. At this stage, reviewers determine final eligibility.

Title/abstract screening is one of the most labor-intensive tasks in the review process. Often, authors face the daunting task of reading through thousands to hundreds of thousands of titles/abstracts. Research suggests that each title/abstract may take anywhere from 30 to 60 seconds to read [28], depending on the expertise and experience of the reviewer. For instance, authors of a SR on the use of nonrestorative treatments for carious lesions screened 9698 titles and abstracts [29]. If a research team of two experienced reviewers reads these in duplicate (i.e., each citation is read independently by two people) at a rate of 30 seconds per title/abstract and dedicates half of their workweek (20 h) to this task, they will finish in about one month. Screening is also prone to human error, as a result of fatigue associated with carrying out this monotonous task.

In this perpetual tension between the need to synthesize information rapidly and the vast amount of time it takes to complete a SR, reviewers (often those with limited resources) may opt for more specific searches in an attempt to reduce the sheer volume of titles and abstracts to screen. These modifications have serious implications and may threaten the validity of the review [11].

Out of all stages in the review process, the integration of ML at the screening stage has reached the highest level of maturity and is ready for end users (i.e., SR authors) [30]. In short, humans have assigned a set of titles/abstracts, using their pre-defined eligibility/selection criteria, in the shape of "Yes/No" decisions about their potential eligibility. This set of references represents

the "training set" and is typically assigned to humans randomly [11, 30]. The machine then learns from these human defined decisions on the "training set" and consequently trains a classifier. Next, the machine makes predictions about the eligibility of previously unseen titles/abstracts using the trained classifier. The titles and abstracts that the machine deems most likely to be eligible (i.e., have the largest probability for inclusion) are reordered to appear first in the list of titles/abstracts. Human then manually screens another batch of titles/abstracts. This time they screen the titles/abstracts (usually about 25 abstracts) [31, 32] that the machine predicted to be most relevant, instead of a random sample. The loop repeats itself, and the classifier model improves with each iteration. The process ends when the reviewer is confident that the machine is adequately trained (Fig. 12.3) [30]. This approach is called "certainty-based active learning" or semi-automation because the human supervises the ML system [28, 30, 33].

To date, there are a handful of fully developed ML tools/software that can assist review authors in completing title/abstract screening more efficiently. These tools require little to no coding background to use, making them attractive and user friendly for SR authors to incorporate in their workflow. Some are freely available [31, 33–37], while others are associated with a fee [32, 38, 39]. The underlying premise of the tools is similar, but each uses a proprietary ML technique or model. The developers of these tools report a variety of metrics to measure their performance. The most commonly reported metrics are recall or sensitivity (i.e., maximizing the identification of all relevant studies) and precision or specificity (i.e., minimizing the inclusion of irrelevant studies). Other tools report "work saved oversampling," to quantify how much work is saved by using ML over manual screening [11]. It is difficult to compare the performance of the available tools against one another because of the variety of metrics reported in the literature. Additionally, no one has compared the performance of the tools against

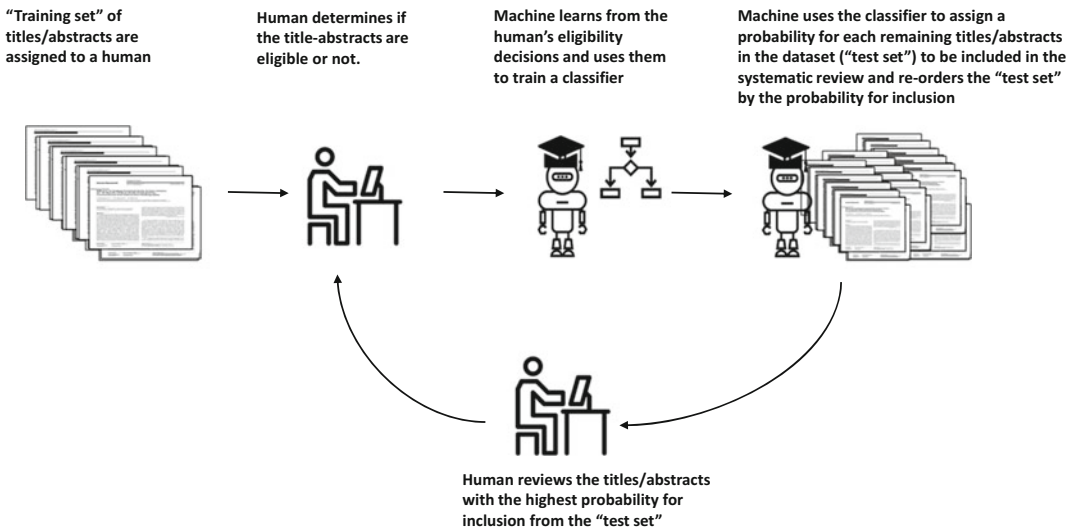


Fig. 12.3 Machine learning in title and abstract screening. (Adapted from: Marshall, I.J., Wallace, B.C., 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 8, 163. https://doi.org/10.1186/s13643-019-1074-9). Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated

one another using the same dataset, which we expect, will facilitate users' decisions regarding ML implementation [40].

The use of ML to semi-automate the title/abstract screening process has several practical implications [11]:

1. Replace the second review: Current methodological standard suggests that two reviewers independently read each title/abstract to reduce human error and to ensure the eligibility criteria are interpreted consistently. In the context of ML, the model proposes that the human screened all titles/abstracts, and the machine acts as the second screener. In a hybrid approach, the machine assigns a second human reviewer a substantially smaller list of studies to screen. This strategy may reduce screening time by at least 50% [11].
2. Reduce the number of titles/abstracts to screen: In areas like public health and social science, the quantity of titles/abstracts obtained from searches is often large, because there are no validated study design filters for non-randomized studies [41] and the research questions are often broad and require information from a variety of research inputs [39]. Through active ML, human screeners stop screening when the machine is fully trained. The point or threshold at which the human stop manually screening is subjective and varies by topic. Studies examining ML for this purpose, reported anywhere from 10% to 90% reduction in the number of titles/abstracts manually screened [35, 36, 42].
3. Increase the rate of screening: Active ML can discern novice from expert reviewers on a team and strategically assign references to each depending on the time it will take to annotate an abstract and its relevance [33]. This approach to screening is known as "efficient citation assignment" [11].
4. Prioritize the order of references: Active ML can prioritize the review of most relevant references first. This approach will allow for the most relevant articles to be retrieved quickly and extracted if deemed relevant simultaneously while screening proceeds. Also, more expert reviewers on a team can retrieve and read relevant articles quicker while novice reviewers screen the less relevant references. Additionally, by viewing more relevant articles in the beginning, screeners will have a better understanding of what references meet the selection criteria and will be more equipped and less likely to be overinclusive as they reach more ambiguous titles/abstracts. This advantage is similar to the second one we mentioned above, but the focus is more on optimizing the workflow [43–46].

12.5 Data Extraction

Extracting data from the primary studies included in the review is one of the most time-consuming tasks and requires a certain level of training and understanding of clinical research concepts. This task is further complicated when primary researchers fail to adhere to reporting standards (e.g., the Consolidated Standards of Reporting Trials (CONSORT)) [47], a reality extensively documented in dental journals [48–53].

Although researchers conducting reviews utilize a series of sources to inform their research question, the most typical one is the full-text scientific article in PDF format. Data in the context of a SR is defined as "... any information about (or derived from) a study, including details of methods, participants, setting, context, interventions, outcomes, results, publications, and investigators" ([54], p. 5). In addition, researchers need to extract specific features of an included primary study to inform the risk of bias assessment (discussed in the next section). To minimize the chance of errors and bias, for each study, two researchers extract the data independently, using a standardized data extraction form previously piloted for optimal usability [54].

Methods for semi-automating data extraction via ML and natural language processing (often just referred to as "NLP") are rapidly evolving and constitute a small yet promising area of research [55]. Extracting data from abstracts and full-text articles can be framed as a text classification problem. Tokens, which are individual words

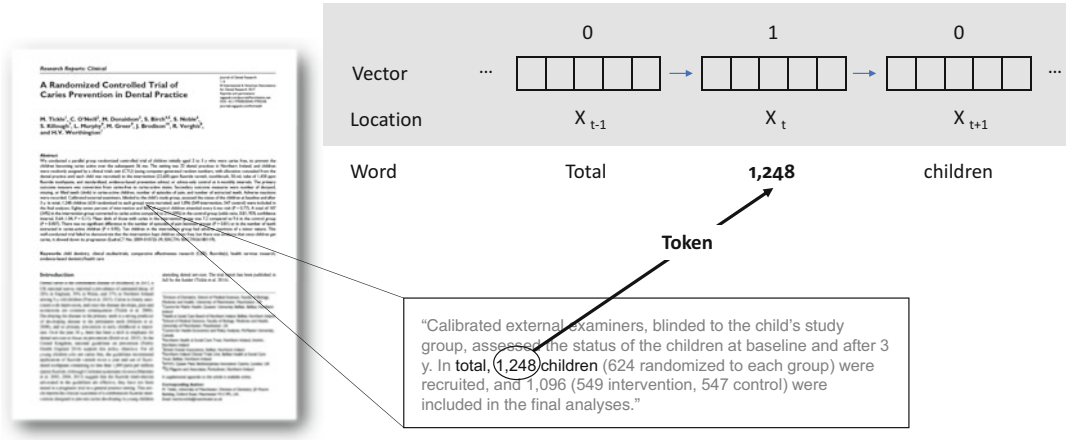


Fig. 12.4 Representation of the use of machine learning for data extraction as a type of text classification problem (e.g., total sample size). (Adapted from: Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Jul11;8(1):163; Abstract from: Tickle M, O'Neill C, Donaldson M, Birch S, Noble S, Killough S, Murphy L, Greer M, Brodison J, Verghis R, Worthington HV. A Randomized Controlled Trial of Caries Prevention in Dental Practice. J Dent Res. 2017 Jul;96(7):741–746). Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated

or numbers representing information elements are classified as relevant. The system then identifies a word potentially relevant to be extracted and creates a vector, where the token or word of interest is tagged, and contextual information is also identified (words in the proximity of the token that contribute to predicting that the word or number is relevant for an information element). While in the vector, the token is in position *t,* the words before and after the token are in position $t - 1$ and $t + 1$, respectively [30]. For example, when identifying the information element, total sample size, a number is the token (i.e., 1248), while the accompanying words "total" ($t - 1$) and "children" ($t + 1$) provide context for the prediction (Fig. 12.4).

For SRs of randomized controlled trials, there are two applications available: RobotReviewer and ExaCT. In general terms, the training data for these two systems utilize full-text articles with sentences manually labelled for relevance. Then, the system retrieves several potential sentences from the full-text PDF (main text and abstract) that may inform a particular information element. Finally, the reviewer is presented with a collection of candidate sentences and chooses the most relevant for an element. Although reviewers looking at a few candidate sentences would be following a more efficient process compared to dealing with the entire PDF, these platforms are categorized as semi-automatic, as they still require human input for final ascertainment [30].

The ExaCT system is a web application that extracts sentences from full-text articles, which are converted into HTML/XML or plain text format to populate a pre-defined template with more than 21 key characteristics of randomized controlled trials (Table 12.2). First, the tool identifies the heading and subheadings of the articles' sections. Then, it detects sentences nested within the sections and classifies them using Support Vector Machine algorithms. Using the bag-of-words approach, the algorithm classifies each sentence according to the probability of informing a specific information element. Finally, ExaCT provides a ranking with the top five sentences showing the

Table 12.2 Set of information elements (key characteristics) of randomized controlled trials available for extraction in the ExaCT system

Element	Description
Eligibility criteria	Logical conditions for being included in the trial, usually split into inclusion and exclusion criteria
Sample size	The total number of participants actually enrolled (randomized) in the trial
Start date of enrolment	Date the enrolment actually started, including day, month, year, or as much as presented
End date of enrolment	Date the enrolment actually ended, including day, month, year, or as much as presented
Name of experimental treatment	Name of experimental intervention
Name of control treatment	Name of control intervention
Dose	Dosage of experimental/control intervention
Frequency of treatment	Frequency of administration of experimental/control intervention
Route of treatment	Route of administration of experimental/control intervention
Duration of treatment	Duration of administration of experimental/control intervention
Primary outcome name	The outcome(s) of greatest importance, where outcome is a "component of a participant's clinical and functional status after an intervention has been applied, that is used to assess the effectiveness of an intervention" (source: Glossary of terms in the Cochrane Collaboration)
Primary outcome time point	Point in time when a primary outcome was assessed
Secondary outcome name	Outcome(s) used to evaluate additional effects of the intervention deemed a priori as being less important than the primary outcomes (Source: Glossary of terms in the Cochrane Collaboration)
Secondary outcome time point	Point in time when a secondary outcome was assessed
Funding organization name	Name of a funding source
Funding number	Funding grant number
Early stopping	Whether the trial was stopped earlier
Registration identifier of trial	Trial registration ID, often ClinicalTrials.gov NCT number
Author name	First and last name of the first author
Date of publication	Year the article was published
DOI	Digital object identifier for the publication

From Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak 2010; 10:56

highest chance for containing the relevant information. Developers of the application describe that the system is further enhanced by using a hierarchy of elements to contextualize the data extracted from the original study and increasing the final prediction of the information element [56]. We will describe RobotReviewer in the risk of bias section of this chapter.

12.6 Appraisal: Risk of Bias Assessment

Evaluating the trustworthiness of the available evidence is a crucial feature of SRs. Empirical evidence suggests that systematic error or bias creates distorted (under-estimation or overestimation of the true effect) estimates of treatment effects or associations [57]. Bias corresponds to a "systematic deviation from the underlying truth because of a feature of the design or conduct of a research study" [58]. One of the most popular tools to assess the potential presence of bias in randomized controlled trials is the Cochrane Risk of Bias tool. This tool measures a construct called "risk of bias," as it is usually impossible to fully ascertain the extent to which an estimate is affected by systematic error, and only the suspicion or potential "risk for bias," more appropriately reflect the intent of the tool [59]. Although the Cochrane Collaboration has recently released a new version of the tool [60], all the available systems using ML to assist reviewers on conducting risk of bias assessment of trials, are based on the 2011 version [61]. The efforts described below are all focusing on such a version of the risk of bias tool.

The 2011 Cochrane Risk of Bias Tool for Randomized Controlled Trials includes six bias domains and sources of bias: (1) selection bias (random sequence generation and allocation concealment), (2) performance bias (blinding of participants and personnel), (3) detection bias (blinding of outcome assessment), (4) attrition bias (incomplete outcome data), (5) reporting bias (selective outcome reporting), and (6) other biases. The judgment about the risk of bias is reflected in three categories of responses for each domain: (1) low risk of bias, (2) high risk of bias, and (3) unclear risk of bias. For transparency, users of the tool are instructed to provide quotations from the original study to support their assessment [61]. The entire evaluation for a single study requires two reviewers acting independently, in a process that could take, on average, approximately 20 minutes per rater [62].

Using ML to assess the risk of bias of randomized controlled trials is a two-step approach. First, it is necessary to identify relevant sentences from the full-text document, and second, assign a risk of bias judgment for each of the domains [63]. Datasets for training usually take advantage of the pre-existing risk of bias assessments contained in the Cochrane Database of Systematic Reviews. These assessments contain a final judgment as to low, high, or unclear risk of bias, and a quote from the full text justifying the answer. High and unclear risk of bias labels are combined to generate a binary variable (i.e., low risk of bias and not low risk of bias) [63]. When using logistic regression to rank sentences (sentence models) according to relevance for each risk of bias domain, a sentence is tagged using one of three labels: relevant, not relevant, or unlabeled. Then datasets are built regrouping the labels depending on the purpose (e.g., relevant/not relevant, relevant/not relevant + unlabeled). These datasets are used to train the parameters for the logistic regression model to determine which one has the best performance. The final steps require using logistic regression to create a ranking to predict the risk of bias at a study level using as input the content from full texts, the title of the article, and the abstract (article model) [63] (Fig. 12.5).

The RobotReviewer system uses a "... multitask linear-kernel Support Vector Machine (SVM) variant; and, (2) A Convolutional Neural Network (RA-CNN). Both models are rationale augmented; that is, they simultaneously classify articles (as being at high/unclear or low risk of bias) and identify text snippets in the article that most strongly support this classification" [64, 65]. A recent on-line based randomized trial suggests that the use of the semi-automating RobotReviewer system may reduce the amount of time invested in conducting the task by 25%

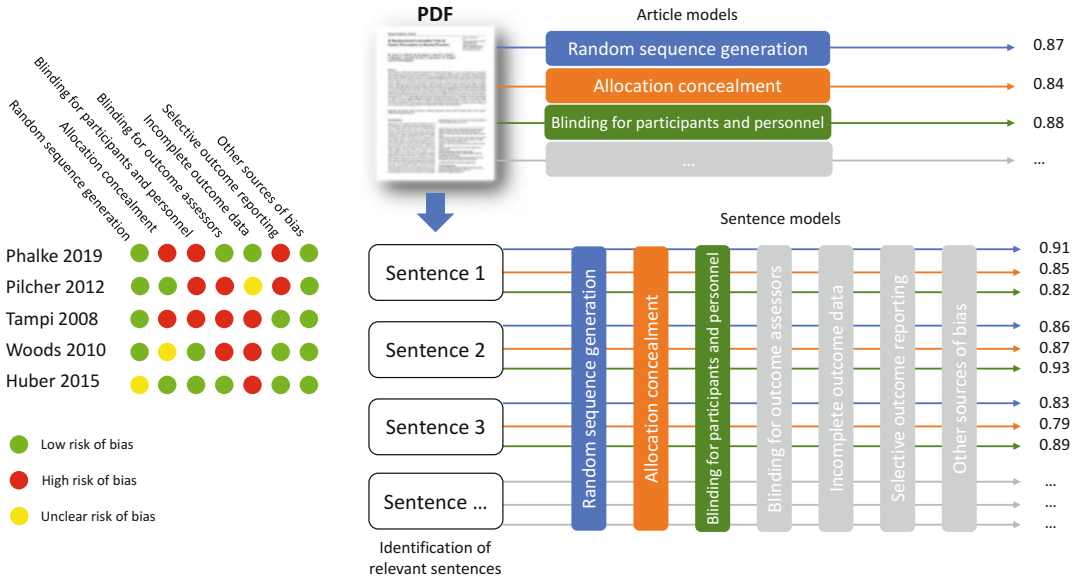


Fig. 12.5 Representation of the use of machine learning for conducting risk of bias assessment of randomized controlled trials. (Adaptation from Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. Int J Epidemiol. 2016 Feb;45(1):266–277) This is an open-access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. You are not required to obtain permission to reuse this article

compared to participants who were allocated to implement a fully manual approach, and provide evidence that the system has high levels of usability, measured by the System Usability Scale [65, 66].

12.7 Implementation Challenges/Barriers to Adoption

Although much progress has been made to automate parts of the SR process, there are several barriers to adopting the existing systems in practice. For one, tools for each distinct task must be interoperable (i.e., data can flow from one system to the next) to be easily integrated into the SR workflow [21]. Additionally, end-users with little knowledge about how the tools function are skeptical of their utility and validity. Some people think ML fully automates the process when in reality, it is semi-automated and human supervised. Developers can change the way they communicate about their systems and use terminology like "computer-assisted" to combat this misconception. Ordinary users also may not think that the ML algorithms are valid. Developers need to demonstrate that their methods are sound and reproducible [40]. Reproducibility can be facilitated by the availability of public datasets [21, 40]. From a developers' perspective, one key limitation of the current tools for data extraction is the limited volume of the available training datasets [30].

The tools developed for each task, also have their own sets of limitations:

Searching

To date, ML has solely been optimized for filtering for RCTs. In the public health realm, observational studies are typical of interest, and such a tool would help narrow down the large pools of research retrieved by searches in this field.

Screening

Screening using ML comes with a host of potential challenges. First, the ML tools described in this chapter were all built for title/abstract screening. Unfortunately, its application in full-text screening is difficult, and no tools exist.

Second, existing tools are not able to identify 100% of all relevant studies (i.e., 100% sensitivity) but, depending on the tools intended function (e.g., replace the second screener), this may be acceptable. It is also important to note that humans do not screen with 100% accuracy, so researchers may be willing to accept a machine that does not do so either in exchange for increased efficiency. Third, the threshold at which the human can stop manual screening is not defined.

Fourth, citations in the training set may not be representative of the "population" of citations; this is known as "hasty generalization." Last, training datasets are imbalanced (contain more excluded than included studies). Imbalanced training sets are problematic because there is less relevant information that the classifier can be trained on. One way to deal with this issue is to "under-sample" excluded studies for the training dataset [11].

Risk of bias

The overall conclusion from the research informing the performance of all these tools is that they still require human supervision before making a final judgment about the risk of bias. However, their ability to identify relevant sentences to inform the domains is promising, as it was evaluated as comparable to the quality of the sentences identified by humans in Cochrane reviews [64].

Using a broader scope, it is also important to acknowledge that any application of ML could widen healthcare inequality. For instance, scientists from the University of California, Berkeley, and the University of Chicago presented evidence of "significant racial bias" in an algorithm that determines health care decisions for more than 70 million people in the United States [67]. While, at the moment, there is no evidence that ML deployed to support SRs has so far introduced such bias, a proactive approach to prevent such bias is certainly desirable. SRs are mostly conducted including licensed health care providers who provide the necessary assurance that sensible results congruent to current biomedical understanding, are published. On the other hand, designers of algorithms for ML that are used to support SRs, may produce software that has a significant impact on public well-being, while operating with relatively limited guidance or oversight [68].

For using ML for SRs, and for any other form of AI for health care, it is important to understand the limitations and to follow the maxim "do noharm" that can perhaps best be upheld by the development of processes and policies to ensure the transparency and reproducibility of AI methods and results [69].

Tools Mentioned in This Chapter

RCT tagger: http://arrowsmith.psych.uic.edu/cgibin/arrowsmith_uic/RCT_Tagger.cgi?ID=15634

Robot Search: https://robotsearch.vortext.systems/

Cochrane CRS: http://crsweb.cochrane.org/login.html

AbstrackR: http://abstrackr.cebm.brown.edu/account/login

SWIFT review: https://www.sciome.com/swift-review/

RobotAnalyst:https://www.google.com/url?q=http://www.nactem.ac.uk/robotanalyst/&sa=D&ust=1570651781470000&usg=AFQjCNELoyBA-49ljcJHWNsUD7BgjRJgHg

GAPscreener: https://omictools.com/gapscreener-tool

EPPI-Reviewer 4: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2935

ExaCT tool: http://exactdemo.iit.nrc.ca

RobotReviewer: https://robot-reviewer.vortext.systems/

References

1. Brignardello-Petersen R, Carrasco-Labra A, Glick M, Guyatt GH, Azarpazhooh A. A practical approach to evidence-based dentistry: understanding and applying the principles of EBD. J Am Dent Assoc. 2014;145:1105–7. https://doi.org/10.14219/jada.2014.102.
2. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312:71–2.
3. Chandler J, Cumpston M, Thomas J, Higgins JP, Deeks JJ, Clarke MJ. Chapter 1: introduction. In: Higgins JPT, Thomas J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated august 2019); 2019. Cochrane.
4. Read K, Creamer A, Kafel D, Vander Hart RJ, Martin ER. Building an escience thesaurus for librarians: a collaboration between the National Network of libraries of medicine, New England region and an associate fellow at the National Library of medicine. J eSci Librariansh. 2013;2:53–67.

5. Tricco AC, Antony J, Zarin W, Strifler L, Ghassemi M, Ivory J, Perrier L, Hutton B, Moher D, Straus SE. A scoping review of rapid review methods. BMC Med. 2015;13:244. https://doi.org/10.1186/s12916-015-0465-6.
6. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Res Synth Methods. 2014;5:371–85. https://doi.org/10.1002/jrsm.1123.
7. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review–a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10(Suppl 1):21–34. https://doi.org/10.1258/1355819054308530.
8. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017;7:e012545. https://doi.org/10.1136/bmjopen-2016-012545.
9. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010;7:e1000326. https://doi.org/10.1371/journal.pmed.1000326.
10. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006;13:206–19. https://doi.org/10.1197/jamia.M1929.
11. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4:5. https://doi.org/10.1186/2046-4053-4-5.
12. Marshall CM. Sutton A. SR tool box [WWW Document]. 2019.
13. Thomas J, Kneale D, McKenzie JE, Brennan SE, Bhaumik S. Chapter 2: determining the scope of the review and the questions it will address. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
14. Kugley S, Wade A, Thomas J, Mahood Q, Klint-Jorgensen AM, Hammerstrom K, Sathe N. A guide to information retrieval for cambell systematic reviews. 2010.
15. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev. 2014;3:74. https://doi.org/10.1186/2046-4053-3-74.
16. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc. 2015;22:707–17. https://doi.org/10.1093/jamia/ocu025.
17. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Res Synth Methods. 2018;9:602–14. https://doi.org/10.1002/jrsm.1287.
18. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. J Am Med Inform Assoc. 2017;24:1165–8. https://doi.org/10.1093/jamia/ocx053.
19. Cochrane Crowd. 2019.
20. Glanville J, Lefebvre C, Wright K. ISSG search filter resource York (UK): the InterTASC information specialists' sub-group [WWW document]. 2019. https://sites.google.com/a/york.ac.uk/issg-search-filters-resource/home. Accessed Sep 10 2019.
21. O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Shemilt I, Thomas J, Glasziou P, Wolfe MS. Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the international collaboration for automation of systematic reviews (ICASR). Syst Rev. 2019;8:57. https://doi.org/10.1186/s13643-019-0975-y.
22. Levay P, Craven J. Systematic searching- chapter 7. London: Facet Publishing; 2019.
23. Introduction what is MeSH? [WWW Document]. 2019. https://www.nlm.nih.gov/bsd/disted/meshtutorial/introduction/index.html Accessed Sep 10 2019.
24. Soto AJ, Przybyla P, Ananiadou S. Thalia: semantic search engine for biomedical abstracts. Bioinformatics. 2018;35:1799–801.
25. Incorporating values for indexing method in MEDLINE/PubMed CML [WWW Document]. 2018.
26. Jimeno-Yepes A, Mork JG, Wilkowski B, Fushman DD, Aronson AR. MEDLINE MeSH indexing: lessons learned from machine learning and future directions. 2012.
27. Mork J, Aronson A, Demner-Fushman D. 12 years on - is the NLM medical text indexer still useful and relevant? J Biomed Semantics. 2017;8:8. https://doi.org/10.1186/s13326-017-0113-5.
28. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinform. 2010b;11:55. https://doi.org/10.1186/1471-2105-11-55.
29. Urquhart O, Tampi MP, Pilcher L, Slayton RL, Araujo MWB, Fontana M, Guzmán-Armstrong S, Nascimento MM, Nový BB, Tinanoff N, Weyant RJ, Wolff MS, Young DA, Zero DT, Brignardello-Petersen R, Banfield L, Parikh A, Joshi G, Carrasco-Labra A. Nonrestorative treatments for caries: systematic review and network meta-analysis. J Dent Res. 2019;98:14–26. https://doi.org/10.1177/0022034518800014.

30. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8:163. https://doi.org/10.1186/s13643-019-1074-9.
31. Przybyła P, Brockmeier AJ, Kontonatsios G, Pogam M-AL, McNaught J, von Elm E, Nolan K, Ananiadou S. Prioritising references for systematic reviews with RobotAnalyst: a user study. Res Synth Methods. 2018;9:470–88. https://doi.org/10.1002/jrsm.1311.
32. Thomas J, Brunton J, Graziosi S. EPPI-reviewer 4: software for research synthesis. London: Social Science Research Unit, UCL Institute of Education; 2010.
33. Wallace BC, Small K, Brodley C, Lau J, Trikalinos TA. Modeling annotation time to reduce workload in Comparative Effectivenss reviews. 2010a.
34. Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, Holmgren S, Pelch KE, Walker V, Rooney AA, Macleod M, Shah RR, Thayer K. SWIFT-review: a text-mining workbench for systematic review. Syst Rev. 2016;5:87. https://doi.org/10.1186/s13643-016-0263-z.
35. Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, Lill CM, Cohen JT, Trikalinos TA. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. Genet Med. 2012a;14:663–9. https://doi.org/10.1038/gim.2012.7.
36. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In Proceedings of the 2nd ACM SIGHIT International Health Informatics. 2012b.
37. Yu W, Clyne M, Dolan SM, Yesupriya A, Wulf A, Liu T, Khoury MJ, Gwinn M. GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC Bioinform. 2008;9:205. https://doi.org/10.1186/1471-2105-9-205.
38. DistillerSR [WWW Document]. Systematic review and literature review software by evidence partners. 2019. https://www.evidencepartners.com/products/distillersr-systematic-review-software/. Accessed Oct 10 2019.
39. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, Kelly MP, Thomas J. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods. 2014;5:31–49. https://doi.org/10.1002/jrsm.1093.
40. Olorisade BK, Brereton P, Andras P. Reproducibility of studies on text mining for citation screening in systematic reviews: evaluation and checklist. J Biomed Inform. 2017;73:1–13. https://doi.org/10.1016/j.jbi.2017.07.010.
41. Reeves BC, Deeks JJ, Higgins JPT, Shea B, Tugwell P, Wells GA. Chapter 24: Including non-randomized studies on intervention effects. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane. 2019.
42. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc. 2010;17:446–53. https://doi.org/10.1136/jamia.2010.004325.
43. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. AMIA Annu Symp Proc. 2008;2008:121–5.
44. Cohen AM, Ambert K, McDonagh M. Studying the potential impact of automated document classification on scheduling a systematic review update. BMC Med Inform Decis Mak. 2012;12:33. https://doi.org/10.1186/1472-6947-12-33.
45. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc. 2009;16:690–704. https://doi.org/10.1197/jamia.M3162.
46. Wallace BC, Small K, Brodley CE, Trikalinos TA. Who should label what? Instance allocation in multiple expert active learning. SDM. 2011. https://doi.org/10.1137/1.9781611972818.16.
47. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med. 2010;152:726–32. https://doi.org/10.7326/0003-4819-152-11-201006010-00232.
48. Al-Namankany AA, Ashley P, Moles DR, Parekh S. Assessment of the quality of reporting of randomized clinical trials in paediatric dentistry journals. Int J Paediatr Dent. 2009;19:318–24. https://doi.org/10.1111/j.1365-263X.2009.00974.x.
49. Hua F, Deng L, Kau CH, Jiang H, He H, Walsh T. Reporting quality of randomized controlled trial abstracts: survey of leading general dental journals. J Am Dent Assoc. 2015;146:669–678.e1. https://doi.org/10.1016/j.adaj.2015.03.020.
50. Kloukos D, Papageorgiou SN, Doulis I, Petridis H, Pandis N. Reporting quality of randomised controlled trials published in prosthodontic and implantology journals. J Oral Rehabil. 2015;42:914–25. https://doi.org/10.1111/joor.12325.
51. Lempesi E, Koletsi D, Fleming PS, Pandis N. The reporting quality of randomized controlled trials in orthodontics. J Evid Based Dent Pract. 2014;14:46–52. https://doi.org/10.1016/j.jebdp.2013.12.001.
52. Sarkis-Onofre R, Poletto-Neto V, Cenci MS, Pereira-Cenci T, Moher D. Impact of the CONSORT statement endorsement in the completeness of reporting of randomized clinical trials in restorative dentistry. J Dent. 2017;58:54–9. https://doi.org/10.1016/j.jdent.2017.01.009.
53. Savithra P, Nagesh LS. Have CONSORT guidelines improved the quality of reporting of randomised controlled trials published in public health dentistry journals? Oral Health Prev Dent. 2013;11:95–103. https://doi.org/10.3290/j.ohpd.a29359.

54. Li T, Higgins JPT, Deeks JJ. Chapter 5: Collecting data. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
55. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. 2015;4:78. https://doi.org/10.1186/s13643-015-0066-7.
56. Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak. 2010;10:56. https://doi.org/10.1186/1472-6947-10-56.
57. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, Als-Nielsen B, Balk EM, Gluud C, Gluud LL, Ioannidis JP, Schulz KF, Beynon R, Welton NJ, Wood L, Moher D, Deeks JJ, Sterne JA. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012;157:429–38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537.
58. Guyatt G, Rennier D, Meade MO, Cook DJ. Glossary. In: Guyatt G, Rennier D, Meade MO, Cook DJ, editors. Users' guide to the medical literature: a manual for evidence-based clinical practice. New York: McGraw-Hill Education; 2015. p. 645.
59. Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
60. Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.
61. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JAC. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;342:d5928.
62. Hartling L, Ospina M, Liang Y. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ. 2009;339:b4012.
63. Millard LAC, Flach PA, Higgins JPT. Machine learning to assist risk-of-bias assessments in systematic reviews. Int J Epidemiol. 2016;45:266–77. https://doi.org/10.1093/ije/dyv306.
64. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc. 2016;23:193–201. https://doi.org/10.1093/jamia/ocv044.
65. Soboczenski F, Trikalinos TA, Kuiper J, Bias RG, Wallace BC, Marshall IJ. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. BMC Med Inform Decis Mak. 2019;19:96. https://doi.org/10.1186/s12911-019-0814-z.
66. Sauro J. Measuring usability with the system usability scale (SUS). 2011. http://www.measuringusability.com/sus.php [WWW document].
67. Nordling L. Mind the gap. Nature. 2019;573:S103–5.
68. Dawson D, Schleiger E, Horton J, McLaughlin J, Robinson C, Quesada G, Scowcroft J, Hajkowicz S. Artificial intelligence: Australia's ethics framework. Data61 CSIRO, Australia. 2019.
69. JASON, The MITRE Corporation. Artificial intelligence for health and health care. JSR-17-Task-002. 2017.

Machine Learning and Deep Learning in Genetics and Genomics

13

Di Wu, Deepti S. Karhade, Malvika Pillai, Min-Zhi Jiang, Le Huang, Gang Li, Hunyong Cho, Jeff Roach, Yun Li, and Kimon Divaris

13.1 Fundamentals of Machine Learning and Oral Health

The notion that neural networks, as universal approximators, could imitate the human brain appeared in literature as early as 1943 [1]. Sixteen years later, the term "machine learning" was coined by Arthur Samuel and has since become an integral tool for assisting clinicians in making medical diagnoses and predicting outcomes

D. Wu (✉)
Division of Oral and Craniofacial Health Science, Adam School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: dwu@unc.edu

D. S. Karhade
Division of Pediatric and Public Health, Adam School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: deepti_karhade@unc.edu

M. Pillai
Carolina Health Informatics Program, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: mpillai@live.unc.edu

M.-Z. Jiang
Department of Applied Physical Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: minzhi@live.unc.edu

L. Huang · Y. Li
Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: lehuang@email.unc.edu; yun_li@med.unc.edu

G. Li
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, USA
e-mail: franklee@live.unc.edu

H. Cho
Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: hunycho@live.unc.edu

J. Roach
Research Computing, University of North Carolina at Chapel Hill, Chapel Hill, USA
e-mail: jeff_roach@unc.edu

K. Divaris
Division of Pediatric and Public Health, Adam School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, USA
e-mail: kimon_divaris@unc.edu

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_13

[2]. Two broad categories of machine learning (ML) methods exist: supervised learning and unsupervised learning. Supervised learning, including classification and regression, learns a model trained on a labeled dataset(s) that maps input features/predictors to the output labels/response/outcome. By contrast, unsupervised learning methods, including clustering, principal component analysis, and density estimation, aim to identify patterns from the data. In a broad sense, three steps allow an ML model to decipher complex patterns from training data and generalize to testing data. First, a model is built using available input data. Next, the model is subsequently tuned and evaluated before it can finally be deployed for the prediction of unknown future outcomes [3]. Deep learning (DL) is a recently developed type of ML, which in many applications can achieve better capacity than traditional ML methods due to many factors including the advances on the hardware end of GPU accelerated technologies and the capability of DL methods to effectively learn latent features. For these reasons, DL has been a faster, if not the most rapidly, developing branch within the ML arena. DL has become so prominent that in recent literature, ML and DL are mentioned mostly as parallel concepts with the former referring only to non-DL ML methods. Following the convention, in this review, we use ML for traditional ML methods such as support vector machines (SVM), random forest, adaptive boosting, and we use DL for methods based on deep neural networks, which are capable of extracting multiple latent features/weights of deep hidden layers (as described in AlexNet [4]) from very large training datasets by harnessing the power of GPUs.

There have been numerous applications of ML in the oral health domain, specifically in dentistry. For instance, machine learning has been used to determine who would be most likely to develop root caries to provide a differential diagnosis for ameloblastoma and odontogenic keratocysts given panoramic radiographs, to classify restorations on panographs and to determine whether children between the ages of 11 and 15 would require extractions prior to orthodontic treatment [5–8]. An ML application has also been reported for early childhood caries (ECC) classification/prediction, where the best performing algorithms were neural networks, followed by SVM [9].

13.2 Machine Learning 'Omics Applications

The choice of computational methods used in a scenario depends on the biological hypotheses to be tested and the data available for testing. Generally speaking, researchers collect data at various molecular levels and from different cellular, subcellular, and intercellular sources. In 1958, Francis Crick proposed the central dogma of molecular biology, which explains the process by which the genetic information in DNA is passed on to messenger RNA (mRNA) to create the functional products: proteins. According to the central dogma, we can appreciate that regulation may occur at the DNA level, the RNA level, the protein level, or at the pathway level involving multiple proteins or complex interplay among DNA, RNA, and proteins. As all of these aspects are relevant, it has become increasingly popular to collect data in genetic, epigenetic, transcriptomic, and proteomics domains, with approaches including genome-wide genotyping arrays, whole genome sequencing, whole exome sequencing, genome-wide methylation arrays, bisulfite sequencing, RNA sequencing (RNA-seq), as well as metabolite and protein quantifications. Of course, environmental factors can also affect molecular function, in domains such as the metabolome and microbiome, and relate to health and disease, but these are beyond the scope of this chapter.

Many statistical methods have been developed to analyze the genetics and genomics data to achieve a better understanding of biomedical questions. In the last few decades, with the development of ML methods such as *k*-means clustering or SVM, we progressively gained the

ability to unearth more information from big data. More recently, given the success of neural network-based representation learning methods in data-driven areas, such as computer vision (CV) and natural language processing (NLP), researchers began applying DL frameworks into genetics and genomics. Due to the similar data structure (e.g., genomic sequence as a sentence) and regular patterns (e.g., genomic sequence motifs as words in one sentence or texture in images), some canonical DL frameworks used in CV and NLP can also be applied to genetic and genomic data analysis.

There are several differences between traditionally used statistical methods and ML/DL. Results from more traditionally used statistical methods may be easier to interpret and thereby explain biological mechanisms. However, these methods heavily rely on assumptions about data distribution, such as linear relationships between variables or specific types of dependence or independence among samples. ML/DL methods have less stringent assumptions regarding data distribution and sample relationships. Thus, they are better equipped to handle high-dimensional, complicated data structures with large sample sizes.

ML has profoundly impacted the interrogation of large, complex 'omics datasets. For example, ML has been used to discover novel genetic loci and biomarkers, and to identify and classify patterns in cancer genomics, leading to a better understanding of the disease process. A recent systematic review examined whether ML could play a prognostic role in the prediction of head and neck cancer [10]. The average reported accuracy of ML techniques in the seven included investigations ranged from 57% to 99%, demonstrating that ML has a promising role in the prognostic prediction of head and neck cancer [10, 11]. Theoretically, integrative systems genomics approaches are more comprehensive and powerful for the detection of genotype–phenotype associations compared to analyses that utilize only one data layer. However, integrative analysis of the multi-omics data is challenging due to the high dimensionality of the data and the complexity of the biological/clinical questions. For these reasons, optimal feature selection and identification of latent patterns in the data are key advantages that ML can offer. If and when successful, integrative 'omics can help investigators uncover and understand the complexity of biological systems and disease processes, identify key molecular features that may operate at different 'omics levels, and ultimately understand how interactions between these elements influence disease risk or other health endpoints. Moreover, multiple associations, emanating from different omics layers, that point towards the same biological pathways are less likely to be false positives [12].

13.3 Multi-Omics and Integrative Analyses

Applications of ML in multi-omics data structures are becoming increasingly common. Technological advancements in the past two decades have facilitated advancements in virtually all 'omics technologies, enabling the generation and analysis of human and microbial DNA (genomics and metagenomics), RNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). In principle, the interrogation of genomic variation can provide information regarding single nucleotide polymorphisms, rare variants, copy number variants, and genomic rearrangements, among others [12]. Transcriptomic analyses generate data on gene expression, alternative splicing long noncoding, small RNAs, etc. The metabolome entails metabolite profiling in various tissues including serum, plasma, urine, cerebrospinal fluid (CSF), and saliva. The proteome contains protein expression-related information such as protein expression levels and posttranslational modifications [12]. These next-generation sequencing and high-throughput 'omics methods produce a wealth of data that, as mentioned previously, are difficult to handle and interpret using traditional statistical methods. Unsuper-

vised ML methods circumvent some of these issues and enable the detection of previously unknown or undetectable patterns in the data, beyond the basic identification of key features by clustering [11]. A prime example application of unsupervised ML in dental research has been the analysis of disease-affected gingival tissues from patients with chronic or aggressive periodontitis. The algorithm by Kebschull and Papapanou [11] demonstrated high diagnostic accuracy when differentiating between chronic and aggressive periodontitis diagnoses based on transcriptome data. Subsequently, unsupervised clustering was utilized with the ultimate goal of detecting novel, pathobiology-based classification of periodontitis [11].

There are two primary methods for 'omics data integration: multi-staged analysis and meta-dimensional analysis. Several statistical tools and software solutions are available for implementing them [13]. Multi-staged analyses separate the data analysis into several steps. In each step, signals are enriched and ultimately the approach enables the discovery of associations both between data types and in relation to the phenotype of interest [12]. Ritchie et al. [12] have described emerging data integration methods to uncover the complex relationships between genomic variation and human phenotypes in a recent review. Examples include genomic variation analysis and domain knowledge-guided analysis. With meta-dimensional analysis, on the other hand, all data types are combined in simultaneous analyses. There are three main subtypes of meta-dimensional analysis: concatenation-based integration, transformation-based integration, and model-based integration [12]. The ultimate goal of utilizing ML and embedded biostatistical analyses to discover genomic and other 'omics biomarkers is to tailor healthcare, including the practice of medicine and dentistry, according to an individual's genetic make-up. Precision medicine is increasingly becoming a cornerstone of medical practice and is no longer a distant vision for the future. For example, an ML algorithm using multi-omics data signatures was able to discriminate between patients with high and low breast cancer risk with over 90% confidence [14]. Precision medicine in dentistry, however, is still in its infancy [15].

13.4 Multi-Omics in Applications in Dental Caries and Early Childhood Oral Health

Dental caries is a common, complex disease that is driven by microbial dysbiosis at the tooth surface–biofilm interface which, if unaltered, can lead to tooth structure dissolution and breakdown [16, 17]. Social, environmental and behavioral factors are known, key influences in the development of dental caries. However, our understanding of the pivotal, causal role of the oral microbiome has rapidly evolved during the last decades, owing largely to technological advances in sequencing methods [18]. It is now well understood that dental caries are characterized by complex metabolic imbalances within the dental biofilm. While the supragingival biofilm maintains an equilibrium of demineralization and remineralization with the underlying tooth structure in health [19], this balance is disrupted in dysbiosis-driven disease. Therefore, a comprehensive understanding of the disease process requires the measurement of these complex biological phenomena at the biofilm–tooth surface interface. Apart from the various 'omics layers that characterize the dysbiotic biofilm environment (i.e., metagenomics, metatranscriptomics, proteomics, and metabolomics) [20], the superimposition and joint analytical consideration of human genome information, a measure of individual susceptibility, is an ambitious but arguably feasible goal. Additional complexity is unavoidable if cross-kingdom interactions are considered in the context of the oral biofilm, via analyses of fungi and viruses [21, 22].

Here we provide an example of a large, community-based, genetic epidemiologic study of early childhood oral health (ZOE 2.0), that

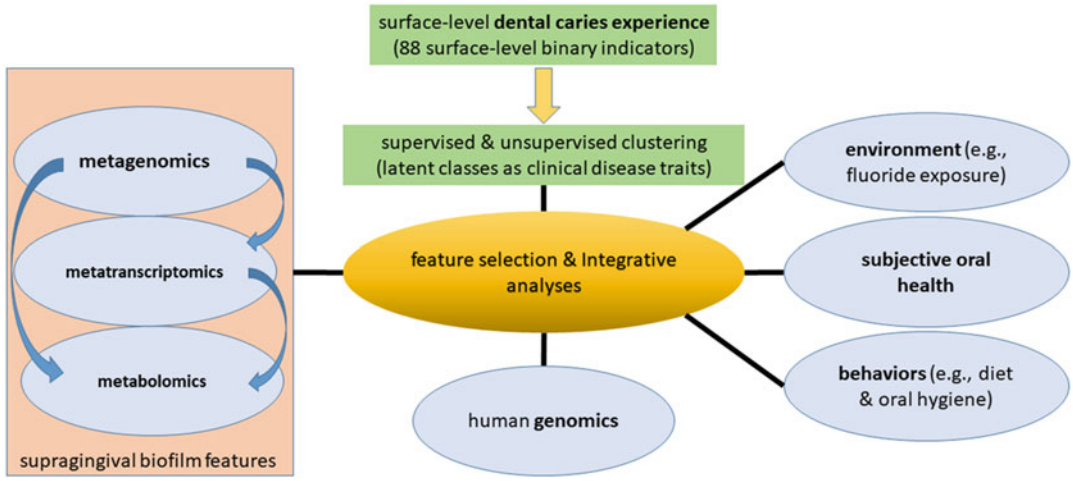


Fig. 13.1 The 'omics, clinical, and behavioral data structure in the ZOE 2.0 early childhood oral health study as an example of an optimally served context for ML applications

has generated four levels of oral health-related omics data: human genomics, supragingival biofilm metagenomics, metatranscriptomics, and metabolomics (from both host and microbiome) [23–25]. Clinical examination data on dental caries experience at the tooth surface level have been collected for approximately 6500 children ages three to five, as well as a wide array of demographic, social, behavioral, and subjective oral health information. Although it has been suggested that the dental caries experience presents in distinct clusters or patterns of clinical presentations [17, 26, 27], this notion has not been widely replicated or formalized in a consensus classification system [28].

As detailed above, the ZOE 2.0 data structure presents both opportunities and challenges (Fig. 13.1). The former revolves around the rich and high-quality data source among a large and well-characterized cohort of young children. The latter is related to the meaningful integration of these multidimensional data structures from a community-based study to answer biological and clinical questions that are aligned with the notion of precision health [15, 29]. ML applications apply to virtually all stages of integration and feature selection. For instance, ML-based methods are applicable at the stage of sequencing classification/alignment in metagenomics analyses, such as MEGAN and PhymmBL among others [30–32]. ML applications are also well-suited for metabolomics and DNA/RNA integrative analyses. Importantly, ML can be used to identify hidden data structures and distinct clusters of clinically-manifest disease departing from a large number of surface-level disease indicators. Further, it can be used to select features of importance from long lists of candidate disease predictors (e.g., environment, subjective measures, behaviors) to aid in the development of predictive models and ultimate decision-making adjuncts. Newer techniques such as AutoML Tables deployed on Google Cloud, which is designed to cover the pipeline from the raw dataset to the deployable ML model, are currently being used by our team to construct ECC classifiers with different sets of context and application-specific predictors and develop predictive models.

Now that we have provided an overview of ML in 'omics data analysis, we will begin to introduce applications of ML/DL in genetics and genomics

in the following sections. The majority of ML methods we discuss will fall under the category of supervised ML.

13.5 Statistical Methods Including Machine/Deep Learning in Association Analysis Between SNPs and Complex Diseases

Common complex human diseases, in contrast to Mendelian diseases, are caused and/or influenced by a combination of genetic and environmental factors. Dental caries, periodontal disease, Alzheimer's disease, scleroderma, asthma, Parkinson's disease, kidney disease, and almost all cancers are examples of complex diseases [33]. With the completion of the Human Genome Project (HGP) [34] researchers have been empowered to investigate the genetic pathogenesis of complex diseases. HGP and subsequent international efforts including the International HapMap Project [35–38] and the 1000 Genomes Project [39–41] catalog common genetic variants segregating among major continental groups and patterns of linkage disequilibrium (LD) among them. These efforts allow for economical and effective mapping of complex diseases, resulting in an improved understanding of the genetic pathogenesis underlying these diseases. Directly benefiting from these international efforts, GWAS have been successfully leveraged to identify genetic variants associated with complex traits and diseases, resulting in over 100,000 associations [42]. Depending on the data type of the trait variable, logistic regression or linear regression can be performed. To detect multi-locus SNP interactions in GWAS, Zhang and Liu [43] proposed an MCMC based method, Bayesian epistasis association mapping (BEAM). Fast epistatic interaction detection using Markov blanket FEPI-MB [44] is a novel and fast Markov Blanket method used to detect SNP interactions by a heuristic search.

Research interests have further expanded from single-locus tests to interaction tests intended to find interacting SNPs jointly associated with a given trait using analysis methods including ML and DL methods.

Two very commonly used ML methods are random forest (RF) and SVM. An early RF-based method required a marginal effect of at least one of the SNP pairs in the SNP pair interaction detection [45]. Several approaches have been developed to address this limitation of RF [46–52]. In this application, SVM outperformed other methods in detecting SNP interactions in both simulated and real data. Furthermore, the results of SVM are easy to interpret and the technique reduces instability in response variables. Nonetheless, some challenges of the applications of SVM include potentially unacceptable type I errors and its applicability to a small number of SNPs. SVM approaches are also fundamentally distance-based and therefore must be extended to properly handle missing data and genetic heterogeneity [53–58]. Although RF techniques can be viewed in the context of a distance-based method [59], the RF framework allows these considerations to be more easily integrated.

DL methods today are normally referred to as neural networks (NNs) based methods. When investigating complex diseases, NNs are commonly trained using known SNPs as inputs and disease traits as outputs. Linkage (identifying linkage between a disease locus and a marker) and association (identifying linkage disequilibrium of a disease locus and a marker) analysis are the two different methods used to discover SNPs associated with a disease [60]. However, DL is not commonly used in association studies of complex diseases.

13.6 ML and DL in CNV Calling

CNVs can be called in normal tissue or cancer tissue. For both, a reference sample is often needed. For example, an adjacent normal tissue sample can be utilized as a reference sample for CNV calling in a tumor sample. Expectation maximization (EM) is a popular

method deployed in various analysis methods [61–64]. ReadDepth [65] is another approach that uses a circular binary segmentation algorithm. Several new methods have been published since 2015 [66–70] including CODEX [71], which involves Poisson likelihood-based recursive segmentation procedures. ML [72–75] and DL [76–78] methods for CNVs are very recent and have yet to be established. Convolutional neural network (CNN)-based method DL-CNV [78], for example, is a novel alignment-free method that calls CNVs from DNA sequencing data of a single sample. In the cancer field, some ML methods have been explored to call both SNVs and CNVs. For example, both Canopy [79], which uses MCMC, and THEMIS [80], which uses dynamic graphical models, can capture SNV and CNV.

Consensus clustering is commonly used for tumor subtype clustering [66]. DL methods have recently been developed for tumor subtype classification. For instance, Karim et al. [67] trained Conv-LSTM and convolutional autoencoder (CAE) to predict cancer types with 80% accuracy. Alshibli, et al. [68] proposed three DL models to classify cancer types based on CNVs: a six-layer convolutional net (CNN6), residual six-layer convolutional net (ResCNN6), and transfer learning of pertained VGG16 net. All these approaches can be applied to study oral cancer.

13.7 ML and DL Methods for DNA Methylation Data

As one of the most commonly used epigenetics measures, DNA methylation plays an important role in the establishment of tissue-specific gene expression and regulation processes. Both high-density array and bisulfite sequencing-based technologies [69] allow the measurement of genome-wide DNA methylation at a high resolution within large population samples.

CpG islands (CGIs) are regions of the genome that contain a large number of CpG dinucleotide repeats, where Cytosine and Guanine are linked by a phosphate in a repeated sequence. These are genetic hotspots as they are sites for active methylation, which may affect the expression of a gene.

With genome-wide methylation data, a plethora of computational tools have been employed for association testing to identify differentially methylated regions (DMRs) associated with phenotypes/traits in EWAS, and to identify genetic variants associated with DNA methylation patterns in methylation quantitative trait loci (mQTL) studies [70].

Recent advances in ML and DL methods have provided us with insights into epigenetic processes and hold the promise for future research as well as clinical and medical applications [81]. On one hand, ML/DL algorithms can be used to impute or predict unobserved DNA methylation states, as a response variable [82–85] or to find DMRs [81]. For example, DeepCpG has been shown to have interpretable model parameters, which can be used to explain how sequence composition affects methylation variability [83]. For another example, Zeng et al. introduced CpGenie [86], a sequence-based DL framework, which learns a regulatory code of DNA methylation through a three-layer CNN to predict the impact of noncoding variants and to produce mQTL.

On the other hand, ML/DL algorithms can utilize methylation data as explanatory features for prediction or classification [86–89]. For example, DNA methylation is used for the classification of central nervous system tumors across all entities and age groups by leveraging the traditional ML method, RF [87]. Cai et al. [88]. utilized DNA methylation data to classify lung cancer using ensemble-based feature selection and ML methods. The algorithm is shown to improve the lung cancer classification accuracy, attaining accuracy of 86.54%, based on leave-one-out cross-validation. Also, DNA methylation of distal regulatory sites was shown to characterize the dysregulation of cancer genes. Constructing the association model between methylation and expression

allowed discrimination of the true methylation-expression pairs and false, randomized pairs [90].

13.8 ML and DL Methods for Hi-C Data

Hi-C is a type of unbiased genome-wide chromosome conformation capture technologies used to characterize chromatin spatial organization inside the nucleus, which helps researchers to understand the function of regulatory elements during many key biological processes including differentiation, meiosis, and division [91]. However, these unbiased and untargeted Hi-C technologies, while having the advantages to capture all-to-all interactions, entail the generation and analysis of unprecedented big data in the order of billions-of-reads per sample. For example, Rao et al. [92] generated 4.9 billion raw reads for the GM12878 lymphoid cell line, Bonev et al. [93] produced 7 billion reads for the mouse embryonic stem cell line, and Jin et al. [94] examined 3.4 billion reads to study IMR90, a primary human fetal lung fibroblast cell line. Although largely needed to attain Kb resolution mapping of chromatin spatial organizations, such billions-of-reads per sample remain cost-prohibitive, along with other challenges associated with the generation of such data including a large number of input materials (particularly for primary tissues), high library preparation costs, and batch effects in the presence of multiple libraries/replicates. To alleviate these issues via in silico methods, several DL methods have been developed recently to enhance the resolution of Hi-C data. For reasons aforementioned, most Hi-C datasets generated to date, with 500 million or fewer reads per sample, have been analyzed at relatively low resolutions (e.g., 25 kb or 40 kb bins). Such coarse resolutions easily harbor multiple genes, multiple genetic variants, or multiple regulatory elements, rendering it challenging to identify more refined structures such as subdomains or enhancer–promotor interactions [95] or to link GWAS variants and distal regulatory elements to their target genes. HiCPlus [95] was first put forward in an attempt to solve the low-resolution problem in Hi-C data. It leverages a DL method to enhance the resolution of a Hi-C contact frequency matrix. The architecture behind HiCPlus's DL method is rather simple and straightforward; it contains three layers of CNN. Albeit simple, the trained model can be utilized to enhance datasets from cell lines other than the training dataset. More recently, to achieve better performance than HiCPlus, particularly to enhance the resolution of low-resolution Hi-C data, HiCNN was proposed by replacing plain networks with the shortcut connections featured by residual networks for more effective gradient propagation in optimization, that have 25-layer of residual blocks, two layers of convolutions in each residual block, and an identity shortcut in each residual block [96]. At the same time, for the same purpose as HiCNN, hicGAN used generative adversarial networks (GAN) to enhance the resolution of HiC data [97].

Computational methods have also been exploited to remove various known and unknown biases (e.g., GC content, mappability, and effective length) from high-resolution Hi-C data [98]. On one hand, for known sources of biases, Yaffe and Tanay [99] proposed a nonparametric method that modeled the probability of observing a paired-end read spanning two fragment ends [99]. In 2012, Hu et al. [100] proposed HiCNorm, which fits a Poisson regression model to remove systematic biases in the raw Hi-C contact maps. With the parametric modeling approach and fewer parameters to be estimated, HiCNorm demonstrated both statistical and computational advantages over the nonparametric method proposed by Yaffe and Tanay [99]. On the other hand, unknown sources of biases can also be accounted for. For example, an iterative correction method [101] based on the Sinkhorn-Knoppbalancing algorithm was proposed to balance the contact frequency matrix. More specifically, the method adopts the following two-step approach to normalize Hi-C data: (1) divide each row by the mean of the rows; (2) divide each column by the mean of the columns. The two steps are repeated until the means converge.

Importantly, computational methods have also played indispensable roles in identifying chromatin interactions from Hi-C data. After bias

correction, researchers typically perform analysis to generate various biologically meaningful readouts from Hi-C data [102, 103], including multi-Mb resolution "A" (i.e., open and actively transcribed) and "B" (i.e., compacted and repressed) compartments, Mb resolution topologically associating domains (TADs), tens-kb resolution Frequently Interacting REgions (FIREs) [104, 105] and high (e.g., kb) resolution chromatin loops [106] and interactions [107–111]. Focusing on the detection of high-resolution chromatin interactions, several statistical methods have been proposed. These include Fit-HiC [112], FitHiC2 [107], HMRF-Bayes [110], and FastHiC [111]. Fit-HiC uses nonparametric splines to estimate mid-range intra-chromosomal contacts. HMRF-Bayes and FastHiC leverage hidden Markov random field-based models to borrow information from neighboring pairs of genomic loci for more robust and powerful detection of chromatin interactions. FastHiC improves upon the Bayesian implementation in HMRF-Bayes by employing the simulated field algorithm to approximate the joint distribution of hidden peak statuses. FastHiC thus efficiently detects biological meaningful chromatin interactions and achieves better performance than Fit-HiC and HMRF-Bayes.

13.9 ML and DL Methods in Analysis of Transcription Factors

Transcription factors (TFs) are proteins that bind to specific genomic regions/sequences to regulate the expression of the gene(s). Understanding the mechanism through which TFs regulate gene expression has been a daunting task. Among many challenges, two tasks remain under the spotlight: (1) identifying TF binding sites (TFBS) and (2) identifying variants that disrupt TF binding.

For the first task, TF motif discovery has been a long-standing problem of keen interest, for which multiple probabilistic approaches have been proposed. For example, an EM algorithm-based method was the first to discover motifs in biopolymers [113]. In the context of TF motif discovery, the focus of inference is to determine where the motif starts and ends in each input sequence. Another approach, Multiple EM for Motif Elicitation (MEME) [114], using a position-specific scoring matrix (PSSM) as input, discovers TF motifs from the input sequences. Various extensions of the MEME method [114] have been proposed, many of which harness information germane to evolutionary conservation to improve the accuracy of predicted motifs. These methods include EmnEM [115], orthoMEME [116], PhyME [117], and Converge [117]. In more recent years, several ML/DL methods have been proposed for TF motif discovery, harnessing the power of concurrent advances in ML, DL, and other artificial intelligence approaches. For example, Alipanahi et al. [118] presented DeepBind to predict sequence specificities, using one layer of CNN to compete favorably with state-of-the-art methods such as MEME-ChIP [119] and MatrixREDUCE [120]. Most recently, Quang et al. [121] proposed FactorNet using a convolutional-recurrent neural network model to predict sequence specificity and showed FactorNet outperformed DeepBind.

Regarding the second task, using computational methods to detect variants that affect TF binding has been of burgeoning interest to many investigators. One paramount motivation is to aid the interpretation of GWAS variants. Over 80% of disease-associated variants identified from GWAS reside in noncoding genomic regions. For the vast majority of these noncoding GWAS variants, the molecular and biological mechanisms remain largely elusive. Among the many potential reasons underlying their associations with disease, an important one is that genetic variants can affect TF binding. Allelic variations at TF binding sites may lead to dysregulated gene expression and ultimately contribute to disease. Many methods have been proposed for this important task. For example, Zhou et al. [122] developed DeepSEA, a DL-based sequence analyzer using CNN blocks to predict the functional effects of noncoding variants. Specifically, the method predicts a

comprehensive epigenetic state of a sequence, including TF binding, DNase I sensitivities, and histone marks in multiple cell types. Despite having only one single layer of CNN, including convolution, Rectified Linear Unit (ReLU) activation function, and max-pooling operation, DeepSEA outperforms a classic RF method, GWAVA [123]. DeFine [124] adopted a CNN approach to predict the impact of noncoding variants on TF binding intensities, which achieved better performance than DeepSEA.

13.10 ML and DL Methods for Single-Cell Transcriptomics Data

With recent developments in single-cell RNA-sequencing (scRNA-seq), geneticists can profile scalable gene expression at the single-cell level, which allows high-resolution dissection of complex biological systems [125–128]. However, the analysis of scRNA-seq data can be challenging due to many confounding factors such as batch effects, variable sensitivity, and gene "dropout" (i.e., zero read-out from sequencing due to a failure to capture RNA for technical reasons) [125–128]. Various ML methods have been proposed to address the above issues. Many of these methods produce denoised scRNA-seq data that can be used for various tasks such as differential expression, clustering, imputation, pseudotime trajectory inference, and construction of gene regulatory networks. Recently, as DL has achieved tremendous successes in various fields (e.g., CV, NLP, drug discovery) [129, 130], many DL methods have also penetrated scRNA-seq data analysis and have already started to deliver on the promise to achieve improved performance for many tasks. Here, we only present several state-of-art DL-based methods for the analysis of scRNA-seq data.

Recently, Deng et al. [131] presented scScope, a scalable deep recurrent approach that can accurately and rapidly identify cell-type composition from millions of single cells. Lopez et al. [132]. introduced single cell variational inference (scVI), a scalable DL framework to learn the probabilistic representation of scRNA-seq data. ScVI appears to be comparable to or outperform existing DL methods like MAGIC [133] and DCA [134], an autoencoder-based DL framework, on different tasks, including clustering and differential expression, while accounting for batch effects and limited sensitivity [135]. Some methods were evaluated in simulated data, including simulation of a rare cell population using a GAN [136]. SCANPY [137] is a Python-based scalable ML toolkit for scRNA-seq data analyses, which incorporates many existing methods, such as SEURAT [138], MONOCLE [139], SCDE/PAGODA [140], MAST [141], CELL RANGER [142], and SCRAN [143]. In contrast to the existing R packages underlying many of its contributing methods, SCANPY's python implementation offers an easy interface to many advanced and computationally more efficient DL libraries such as PyTorch, Keras, and TensorFlow. With the advancements of both scRNA-seq technology and DL methods, we expect more enhanced tools, in terms of both statistical and computational efficiency, to become available to geneticists, which will further empower them to decipher information from scRNA-seq datasets.

13.11 Machine and Deep Learning in Genomics-Based Oral Health: Systemic Review of Literature

To gain a better understanding of the use of ML/DL methods with genetic or genomic data to answer questions in oral health, a systematic review was conducted.

13.11.1 Methods

The methodology presented in the *PRISMA Statement* for systematic reviews (Moher et al. 2009) was used to create the search strategy. A comprehensive search was done using the PubMed database for all studies relevant to oral health,

machine or deep learning, and genetics or genomics. The search was conducted on January 27, 2020, using the query:

(((oral[Title/Abstract] OR dental[Title/Abstract] OR "mouth diseases"[MeSH Terms])
AND
("machine learning" [All Fields] OR "deep learning" [All Fields] OR "machine learning" [mh] OR "machine learning" [majr] OR "deep learning" [mh] OR "deep learning" [majr]))
AND
("genomics" [All Fields] OR "genomics" [mh] OR "genomics" [majr] OR "genetics" [All Fields])
NOT
("review" [Publication Type] OR "review literature as topic" [MeSH Terms] OR "review"[All Fields]).

The search strategy and results are shown in Fig. 13.2. Thirty-nine articles were identified with the search. The inclusion criteria for this review were (1) the study had to be focused on oral health and (2) the study had to use an ML/DL method. The exclusion criterion was that the study had to be primary literature describing original research. If a study included oral health in its abstract, the full text of the study was manually reviewed. If a study used machine or deep learning methods, it was included in the review. Twelve of 39 studies did not pertain to oral health. The remaining 27 studies were included for full-text review. In the full-text review process, three studies did not meet the inclusion criteria. One study was not focused on oral health, one was not primary literature, and one did not use machine or deep learning methods. For the remaining 24 studies, data on disease focus, the method used, and research tasks were extracted.

13.11.2 Results

Twenty-four studies met the criteria for the review. The disease focus for each study was examined to understand the domains of oral health

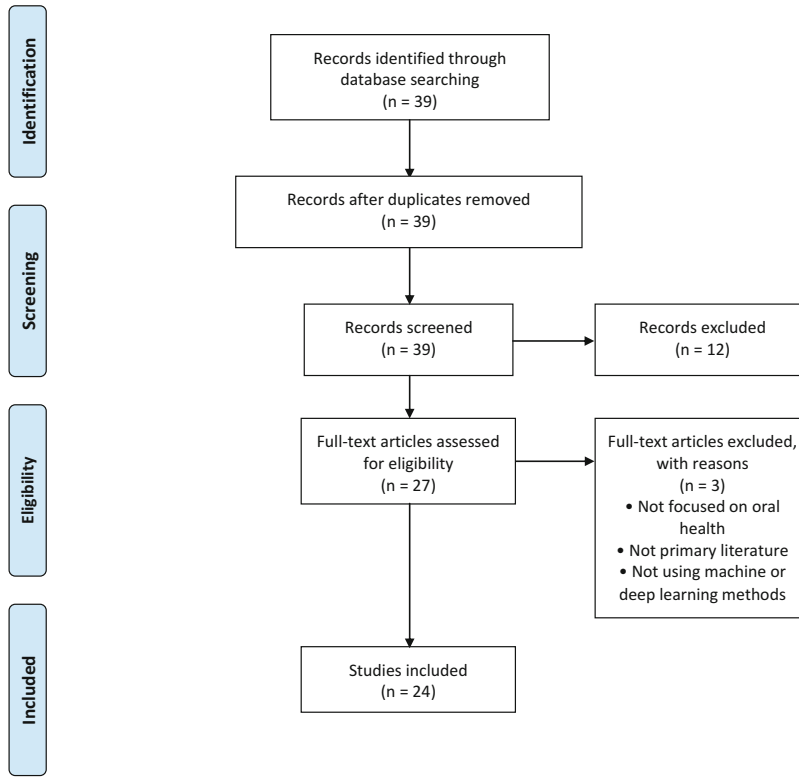


Fig. 13.2 PRISMA flow diagram for systematic review

research. Ten studies pertained to research in cancer, and 14 studies pertained to research in other areas. The other areas and corresponding numbers of studies included periodontal disease (4), periodontitis (2), oral infections (1), oral cleft (2), caries (2), oral microbiota (1), and oral malodor (2).

Studies in the review used the following classification methods: SVM, RF, logistic regression, Naïve Bayes, nearest neighbor clustering, artificial neural network (ANN), decision tree, and deep neural network (DNN). Of all studies included in the review, 8% of studies (2) used deep learning methods as DNN. The studies using DNN were aimed at composition analysis of oral microbiota associated with periodontal disease [144] and predicting malodor [145]. The other 92% of studies (22) used a variety of machine learning methods only. Fifty percent of studies (12) used a single machine learning method in their analyses, while the other half used a range of two to six models for classification to attain optimal performance. The most commonly used machine learning method was SVM, which was used in 79% of studies (19), and the next most common method was RF, which was included in 33% of studies (8). Various studies also used machine learning in feature selection to identify the most discriminating information in each classification task.

The research tasks were then examined to understand the research questions across all studies. The studies were divided into the following research task categories: diagnosis, prediction, diagnosis and prediction, prognosis, and mechanism/biology. Four studies were focused on automated diagnosis. For example, Hsieh et al. [146] presented an approach to separate healthy subjects from patients with oral cavity squamous cell carcinoma using machine learning. Six studies focused on prediction. Nakano et al. [145] predicted oral malodor by examining microbiota from saliva samples using deep learning. Four studies focused on both prediction and diagnosis. Saxena et al. [147] used an SVM to first identify an optimal combination of biomarkers from saliva and pooled plaque samples and consequently used an ANN and decision tree to discriminate between severe early childhood caries and caries-free samples. Three studies focused on disease prognosis. Carnielli et al. [148] used machine learning methods to predict the power of prognostic signatures for oral cavity squamous cell carcinoma. Seven studies focused on investigating underlying biology involved in oral conditions. For example, Torres et al. [149] identified a genome from an oral bacterium in periodontal samples. The various research tasks included a degree of overlapping due to the narrow subject area.

13.12 Discussion of Limitations About Using ML and DL in Genetics and Genomics in Oral Health

ML/DL will play a more important role in prediction and classification to address complicated data structures. However, current dental-related 'omics data studies typically have smaller sample sizes compared to those of other disease fields, which is a limitation of using ML/DL.

It is a ground truth that more flexible models, with fewer fundamental assumptions about the underlying data, may achieve lower bias at the price of higher variance and thus require a larger sample size to overcome the higher variance. ML can be viewed as a toolbox with tools on a spectrum. At the left extreme of the spectrum lies the linear model, and on the right extreme, we find DL methods. The spectrum can be roughly quantified in terms of the Vapnik-Chervonenkis dimension [150]. The methods on the left postulate the most stringent assumption regarding the data and can be considered with minimum sample sizes, although at a price of potentially high bias. Thus, simple models can only be used when the assumed model is believed to be close to the truth.

Some confidence can be given by implementing simple diagnostics. Fortunately, it is not hard to find an example where linear models work in practice. In contrast, ML methods on the right extreme depend on a large number of examples. When larger sample sizes are available, high accuracy can be achieved with

more complex models. For example, the kernel SVM and RF can uncover nonlinear interaction effects that linear models cannot find. CNN's representation learning capacity, which has successfully decreased the classification error rate to less than the human error level, may not be available in most of the non-DL methods.

Thus, when studying 'omics data, knowledge of what methods are available in the ML toolbox, how they are ordered in terms of the trade-off between required sample size and potential bias, and the other advantages and disadvantages of each of them will help researchers maximize the utility of their study design and data.

In this regard, a small sample size with high dimensionality is a challenge in oral health 'omics data. The bias-variance trade-off should be considered with care. Given that the underlying association of a disease with genotypic features can be successfully approximated by simple models, it is worth seriously considering the classical statistical learning methods, such as linear models and generalized linear models possibly with sparsity penalization or after variable screening. Prior knowledge from the domain area can be incorporated into the model by employing the Bayesian framework. In this context, even smaller sample sizes can be handled.

However, ML/DL algorithms can still be considered under relatively small sample sizes. ML is a wide collection of methods ranging from fully nonparametric to fully parametric models. The stronger and more parsimonious a structure is posed, the less sample size is required. For instance, a neural network would require smaller sample sizes if the true model could be approximated by fewer layers and nodes. Contrarily, deep learning, which is composed of an array of multiple layers, may not be directly applicable in these small sample settings without extensive regularization, or even at all for very small sample sizes.

One approach for DL application to small sample data is dimension reduction. Autoencoders, such as deep autoencoder (DAE), sparse autoencoder (SAE), and variational autoencoder (VAE), have been popular nonlinear alternatives to principal component analysis (PCA) [151]. Autoencoders can find a lower-dimensional representation of data in a nonlinear fashion. In this regard, autoencoders are effectively automatically determining the appropriate assumptions required to apply methods suitable for smaller sample sizes. With the lower representation of 'omics data, learning the association between 'omics data and disease status can be done with less accuracy deterioration due to high dimensionality.

This dimension reduction procedure can further be enriched by incorporating external data. While only small numbers of samples of 'omics data are available with the disease status and other phenotypic data, more 'omics data could exist without the disease labels of interest. Although learning a disease prediction model requires samples of all components, dimension reduction, in general, does not require the disease labels (i.e., unsupervised learning) and thereby can entertain the information given by extra or external data.

This lower representation of 'omics data can be fed into a simpler model to learn from small-sized data. DeepMicro [152] is one of the methods that take this DL-based two-step approach in metagenomics. In two-step approaches, intrinsic minimal information is learned about the genomics data without disease labels (i.e., STEP 1, unsupervised learning by autoencoders), and the association between the low-dimensional representation of genomics data and the disease labels are further learned in a supervised way (i.e., STEP 2). In STEP 2, convolutional layers can be constructed to reflect the established knowledge of the phylogenetic tree structure [153].

Ultimately, the number of samples can be increased along with the decreasing sequencing cost.

13.13 Summary and Conclusion of ML and DL in Genetics and Genomics in Oral Health

Statistical and ML methods have been used in many genomics data analysis fields to answer oral health-related questions. The disease phenotypes include, but are not limited to, early childhood

caries, periodontitis, cleft palate, and oral cancer. One of the fundamental challenges is that the disease phenotypes are not always well defined, and the clinical definition may reflect more of the symptoms instead of being associated with biological or genomics mechanisms. Instead of using a single diagnosis value to define a disease, the trend is to use multiple measurements/features to describe the phenotypes. Concurrently, genomics data are typically high-dimensional, have correlation structures, and potentially extensive missing data and zeros. Multiple layers of 'omics data from the same cohort of subjects provide opportunities to investigate diseases more thoroughly, but as the next era of data integration, they also bring challenges.

Here, we introduced an example study from a large cohort in ZOE 2.0 and reviewed ML/DL-based data analysis methods to answer questions in genetics and genomics in biological and clinical fields, including the oral health field. Not all methods discussed have been applied to answer oral health questions, but they are poised to analyze more of the 'omics data collected in oral health studies and dental clinics, particularly using genomics information for precision dental care. In the end, we discussed the reasons and considerations for the limitations of ML/DL in oral genetics and genomics due to the current, relatively small sample size. Generally speaking, the ML/DL tools developed or being developed can be used in answering oral genetics and genomics questions, and hopefully, this will facilitate the oral health clinical diagnosis and decision-making (e.g., oral cancer, caries, and periodontitis), as in some of the other disease fields including breast cancer, lung cancer, among others.

Acknowledgments This work was supported by grants from the National Institutes of Health (NIH), National Institute of Dental and Craniofacial Research, R03-DE028983 to DW and HC, U01-DE025046 to KD and HC, NIH R01 GM105785, R01 HL129132, and R01 HL146500 to YL, and NLM T15-LM012500 to MP.

References

1. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33.
2. Park WJ, Park J-B. History and application of artificial neural networks in dentistry. Eur J Dent. 2018;12(04):594–601.
3. Lin E, Lane H-Y. Machine learning and systems genomics approaches for multi-omics data. Biomarker Res. 2017;5(1):2.
4. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Informn Proc Syst. 2012;25:1097–105.
5. Hung M, Voss MW, Rosales MN, Li W, Su W, Xu J, et al. Application of machine learning for diagnostic prediction of root caries. Gerodontology. 2019;36(4):395–404.
6. Liu Z, Liu J, Zhou Z, Zhang Q, Wu H, Zhai G, et al. Differential diagnosis of ameloblastoma and odontogenic keratocyst by machine learning of panoramic radiographs. Int J Comput Assist Radiol Surg. 2021;16(3):415–22
7. Abdalla-Aslan R, Yeshua T, Kabla D, Leichter I, Nadler C. An artificial intelligence system using machine-learning for automatic detection and classification of dental restorations in panoramic radiography. Oral Surg Oral Med Oral Pathol Oral Radiol. 2020;130(5):593–602.
8. Xie X, Wang L, Wang A. Artificial neural network modeling for deciding if extractions are necessary prior to orthodontic treatment. Angle Orthod. 2010;80(2):262–6.
9. Montenegro RD, Oliveira AL, Cabral GG, Katz CR, Rosenblatt A. A comparative study of machine learning techniques for caries prediction. In: 2008 20th IEEE International Conference on tools with artificial intelligence. Piscataway, NJ: IEEE; 2008. p. 477–81.
10. Patil S, Habib Awan K, Arakeri G, Jayampath Seneviratne C, Muddur N, Malik S, et al. Machine learning and its potential applications to the genomic study of head and neck cancer—a systematic review. J Oral Pathol Med. 2019;48(9):773–9.
11. Kebschull M, Papapanou PN. Exploring genome-wide expression profiles using machine learning techniques. Methods Oral Biol. 2017;1537:347–64. Springer
12. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.

13. Misra BB, Langefeld C, Olivier M, Cox LA. Integrated omics: tools, advances and future approaches. J Mol Endocrinol. 2019;62(1):R21–45.
14. Fröhlich H, Patjoshi S, Yeghiazaryan K, Kehrer C, Kuhn W, Golubnitschaja O. Premenopausal breast cancer: potential clinical utility of a multi-omics based machine learning approach for patient stratification. EPMA J. 2018;9(2):175–86.
15. Divaris K. Fundamentals of precision medicine. Compend Contin Educ Dent. 2017;38(8 Suppl):30–2.
16. Selwitz RH, Ismail AI, Pitts NB. Dental caries. Lancet. 2007;369(9555):51–9. https://doi.org/10.1016/S0140-6736(07)60031-2.
17. Divaris K. Predicting dental caries outcomes in children: a "risky" concept. J Dent Res. 2016;95(3):248–54. https://doi.org/10.1177/0022034515620779.
18. Burne RA, Zeng L, Ahn SJ, Palmer SR, Liu Y, Lefebure T, et al. Progress dissecting the oral microbiome in caries and health. Adv Dent Res. 2012;24(2):77–80. https://doi.org/10.1177/0022034512449462.
19. Marsh PD. Microbial ecology of dental plaque and its significance in health and disease. Adv Dent Res. 1994;8(2):263–71. https://doi.org/10.1177/08959374940080022001.
20. Nyvad B, Crielaard W, Mira A, Takahashi N, Beighton D. Dental caries from a molecular microbiological perspective. Caries Res. 2013;47(2):89–102. https://doi.org/10.1159/000345367.
21. Falsetta ML, Klein MI, Colonne PM, Scott-Anne K, Gregoire S, Pai CH, et al. Symbiotic relationship between Streptococcus mutants and *Candida albicans* synergizes virulence of plaque biofilms in vivo. Infect Immun. 2014;82(5):1968–81. https://doi.org/10.1128/IAI.00087-14.
22. Delisle AL, Guo M, Chalmers NI, Barcak GJ, Rousseau GM, Moineau S. Biology and genome sequence of Streptococcus mutans phage M102AD. Appl Environ Microbiol. 2012;78(7):2264–71. https://doi.org/10.1128/AEM.07726-11.
23. Divaris K, Joshi A. The building blocks of precision oral health in early childhood: the ZOE 2.0 study. J Public Health Dent. 2018;80(Suppl 1):S31–6. https://doi.org/10.1111/jphd.12303.
24. Ginnis J, Ferreira Zandona AG, Slade GD, Cantrell J, Antonio ME, Pahel BT, et al. Measurement of early childhood Oral health for research purposes: dental caries experience and developmental defects of the enamel in the primary dentition. Methods Mol Biol. 1922;2019:511–23. https://doi.org/10.1007/978-1-4939-9012-2_39.
25. Divaris K, Shungin D, Rodriguez-Cortes A, Basta PV, Roach J, Cho H, et al. The Supragingival biofilm in early childhood caries: clinical and laboratory protocols and bioinformatics pipelines supporting metagenomics, Metatranscriptomics, and metabolomics studies of the Oral microbiome. Methods Mol Biol. 1922;2019:525–48. https://doi.org/10.1007/978-1-4939-9012-2_40.
26. Haworth S, Esberg A, Lif Holgerson P, Kuja-Halkola R, Timpson NJ, Magnusson PKE, et al. Heritability of caries scores, trajectories, and disease subtypes. J Dent Res. 2020;99(3):264–70. https://doi.org/10.1177/0022034519897910.
27. Shaffer JR, Feingold E, Wang X, Tcuenco KT, Weeks DE, DeSensi RS, et al. Heritable patterns of tooth decay in the permanent dentition: principal components and factor analyses. BMC Oral Health. 2012;12:7. https://doi.org/10.1186/1472-6831-12-7.
28. GlobalSurg C. Writing g, patient r, statistical a, protocol d, project s, et al. global variation in anastomosis and end colostomy formation following left-sided colorectal resection. BJS Open. 2019;3(3):403–14. https://doi.org/10.1002/bjs5.50138.
29. Divaris K. Searching deep and wide: advances in the molecular understanding of dental caries and periodontal disease. Adv Dent Res. 2019;30(2):40–4. https://doi.org/10.1177/0022034519877387.
30. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46.
31. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17(3):377–86. https://doi.org/10.1101/gr.5969107.
32. Brady A, Salzberg S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods. 2011;8(5):367. https://doi.org/10.1038/nmeth0511-367.
33. Craig J. Complex diseases: research and applications. Nature Education. 2008;1(1):184.
34. The Human Genome Project. https://www.genome.gov/human-genome-project. 2018; Accessed 2020.
35. The International HapMap Consortium. The international HapMap project. Nature. 2003;426(6968):789–96.
36. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299–320.
37. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851–61.
38. The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52–8. https://doi.org/10.1038/nature09298.
39. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73. http://www.nature.com/nature/journal/v467/n7319/abs/nature09534.html#supplementary-information

40. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65. https://doi.org/10.1038/nature11632.
41. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. https://doi.org/10.1038/nature15393.
42. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). Nucleic Acids Res. 2017;45(D1):D896–d901. https://doi.org/10.1093/nar/gkw1133.
43. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nat Genet. 2007;39(9):1167–73.
44. Han B, Chen X-W, Talebizadeh Z. FEPI-MB: identifying SNPs-disease association using a Markov Blanket-based approach. BMC Bioinform. 2011;12(Suppl 12):S3.
45. Uppu S, Krishna A, Gopalan RP. A review on methods for detecting SNP interactions in high-dimensional genomic data. IEEE/ACM Trans Comput Biol Bioinform. 2016;15(2):599–612.
46. Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinform. 2009;10(1):S65.
47. De Lobel L, Geurts P, Baele G, Castro-Giner F, Kogevinas M, Van Steen K. A screening methodology based on random forests to improve the detection of gene–gene interactions. Eur J Hum Genet. 2010;18(10):1127–32.
48. Yoshida M, Koike A. SNPInterForest: a new method for detecting epistatic interactions. BMC Bioinform. 2011;12(1):469.
49. Schwarz DF, König IR, Ziegler A. On safari to random jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics. 2010;26(14):1752–8.
50. Wu Q, Ye Y, Liu Y, Ng MK. SNP selection and classification of genome-wide SNP data using stratified sampling random forests. IEEE Trans Nanobioscience. 2012;11(3):216–27.
51. Lin HY, Ann Chen Y, Tsai YY, Qu X, Tseng TS, Park JY. TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. Ann Hum Genet. 2012;76(1):53–62.
52. Pan Q, Hu T, Malley JD, Andrew AS, Karagas MR, Moore JH. Supervising random forest using attribute interaction networks. European conference on evolutionary computation, machine learning and data mining in bioinformatics. Berlin: Springer; 2013. p. 104–16.
53. Chen SH, Sun J, Dimitrov L, Turner AR, Adams TS, Meyers DA, et al. A support vector machine approach for detecting gene-gene interaction. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2008;32(2):152–67.
54. Özgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24(13):i277–i85.
55. Shen Y, Liu Z, Ott J. Support vector machines with L 1 penalty for detecting gene-gene interactions. Int J Data Min Bioinform. 2012;6(5):463–70.
56. Fang YH, Chiu YF. SVM-based generalized multifactor dimensionality reduction approaches for detecting gene-gene interactions in family studies. Genet Epidemiol. 2012;36(2):88–98.
57. Marvel S, Motsinger-Reif A. Grammatical evolution support vector machines for predicting human genetic disease association. Proceedings of the 14th annual conference companion on Genetic and evolutionary computation 2012. p. 595–8.
58. Zhang H, Wang H, Dai Z, Chen M-S, Yuan Z. Improving accuracy for cancer classification with a new algorithm for genes selection. BMC Bioinform. 2012;13(1):298.
59. Lin Y, Jeon Y. Random forests and adaptive nearest neighbors. J Am Stat Assoc. 2006;101(474):578–90. https://doi.org/10.1198/016214505000001230.
60. Koo CL, Liew MJ, Mohamad MS, Salleh M, Hakim A. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. Biomed Res Int. 2013;2013:432375.
61. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. Bioinformatics. 2016;32(15):2375–7.
62. Ivakhno S, Roller E, Colombo C, Tedder P, Cox AJ. Canvas SPW: calling de novo copy number variants in pedigrees. Bioinformatics. 2018;34(3):516–8.
63. Wang Z, Hormozdiari F, Yang W-Y, Halperin E, Eskin E. CNVeM: copy number variation detection using uncertainty of read mapping. J Comput Biol. 2013;20(3):224–36.
64. Nguyen HT, Merriman TR, Black MA. The CNVrd2 package: measurement of copy number at complex loci using high-throughput sequencing data. Front Genet. 2014;5:248.
65. Miller CA, Hampton O, Coarfa C, Milosavljevic A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS One. 2011;6(1):e16327.
66. Aure MR, Vitelli V, Jernström S, Kumar S, Krohn M, Due EU, et al. Integrative clustering reveals a novel split in the luminal a subtype of breast cancer with impact on outcome. Breast Cancer Res. 2017;19(1):44. https://doi.org/10.1186/s13058-017-0812-y.
67. Karim MR, Rahman A, Jares JB, Decker S, Beyan O. A snapshot neural ensemble method for cancer-type prediction based on copy number variations. Neural Comput & Applic. 2019:1–19.
68. AlShibli A, Mathkour H. A shallow convolutional learning network for classification of cancers based on copy number variations. Sensors. 2019;19(19):4207.
69. Fortin J-P, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina

HumanMethylationEPIC array with minfi. Bioinformatics. 2017;33(4):558–60.
70. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6(8):597–610.
71. Jiang Y, Oldridge DA, Diskin SJ, Zhang NR. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 2015;43(6):e39-e.
72. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665–74.
73. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, et al. QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007;35(6):2013–25.
74. Zhang Z, Cheng H, Hong X, Di Narzo AF, Franzen O, Peng S, et al. EnsembleCNV: an ensemble machine learning algorithm to identify and genotype copy number variation using SNP array data. Nucleic Acids Res. 2019;47(7):e39-e.
75. Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach for accurate detection of copy number variants from exome sequencing. Genome Res. 2019;29(7):1134–43.
76. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7.
77. Hill T, Unckless RL. A deep learning approach for detecting copy number variation in next-generation sequencing data. G3: Genes, Genomes, Genetics. 2019;9(11):3575–82.
78. Zhang Y, Jin L, Wang B, Hu D, Wang L, Li P, et al. DL-CNV: a deep learning method for identifying copy number variations based on next generation target sequencing. Math Biosci Eng: MBE. 2019;17(1):202–15.
79. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci. 2016;113(37):E5528–E37.
80. Liu J, Halloran JT, Bilmes JA, Daza RM, Lee C, Mahen EM, et al. Comprehensive statistical inference of the clonal structure of cancer from multiple biopsies. Sci Rep. 2017;7(1):1–13.
81. Holder LB, Haque MM, Skinner MK. Machine learning for epigenetics and future medical applications. Epigenetics. 2017;12(7):505–14.
82. Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics. 2019;35(22):4586–95.
83. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):67.
84. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16(1):14.
85. Zhang G, Huang KC, Xu Z, Tzeng JY, Conneely KN, Guan W, et al. Across-platform imputation of DNA methylation levels incorporating nonlocal information using penalized functional regression. Genet Epidemiol. 2016;40(4):333–40. https://doi.org/10.1002/gepi.21969.
86. Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA methylation. Nucleic Acids Res. 2017;45(11):e99-e.
87. Capper D, Jones DT, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555(7697):469–74.
88. Cai Z, Xu D, Zhang Q, Zhang J, Ngai S-M, Shao J. Classification of lung cancer using ensemble-based feature selection and machine learning methods. Mol BioSyst. 2015;11(3):791–800.
89. Wei SH, Balch C, Paik HH, Kim Y-S, Baldwin RL, Liyanarachchi S, et al. Prognostic DNA methylation biomarkers in ovarian cancer. Clin Cancer Res. 2006;12(9):2788–94.
90. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14(3):R21.
91. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nat Methods. 2017;14(7):679–85. https://doi.org/10.1038/nmeth.4325.
92. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
93. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, et al. Multiscale 3D genome rewiring during mouse neural development. Cell. 2017;171(3):557–72.e24. https://doi.org/10.1016/j.cell.2017.09.043.
94. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503(7475):290–4. https://doi.org/10.1038/nature12644.
95. Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat Commun. 2018;9(1):750. https://doi.org/10.1038/s41467-018-03113-2.
96. Liu T, Wang Z. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. Bioinformatics. 2019;35(21):4222–8. https://doi.org/10.1093/bioinformatics/btz251.

97. Liu Q, Lv H, Jiang R. hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics. 2019;35(14):i99–i107. https://doi.org/10.1093/bioinformatics/btz317.
98. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75. https://doi.org/10.1016/j.ymeth.2014.10.031.
99. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43(11):1059–65. https://doi.org/10.1038/ng.947.
100. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28(23):3131–3. https://doi.org/10.1093/bioinformatics/bts570.
101. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999–1003. https://doi.org/10.1038/nmeth.2148.
102. Li Y, Hu M, Shen Y. Gene regulation in the 3D genome. Hum Mol Genet. 2018;27(R2):R228–r33. https://doi.org/10.1093/hmg/ddy164.
103. Yu M, Ren B. The three-dimensional Organization of Mammalian Genomes. Annu Rev Cell Dev Biol. 2017;33:265–89. https://doi.org/10.1146/annurev-cellbio-100616-060531.
104. Crowley C, Yang Y, Qiu Y, Hu B, Won H, Ren B, et al. FIREcaller: an R package for detecting frequently interacting regions from Hi-C data. bioRxiv. 2019; 619288. https://doi.org/10.1101/619288.
105. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. 2016;17(8):2042–59. https://doi.org/10.1016/j.celrep.2016.10.061.
106. Rao Suhas SP, Huntley Miriam H, Durand Neva C, Stamenova Elena K, Bochkov Ivan D, Robinson James T, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80. https://doi.org/10.1016/j.cell.2014.11.021.
107. Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc. 2020;15(3):991–1012. https://doi.org/10.1038/s41596-019-0273-0.
108. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014; https://doi.org/10.1101/gr.160374.113.
109. Juric I, Yu M, Abnousi A, Raviram R, Fang R, Zhao Y, et al. MAPS: model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. PLoS Comput Biol. 2019;15(4):e1006982. https://doi.org/10.1371/journal.pcbi.1006982.
110. Xu Z, Zhang G, Jin F, Chen M, Furey TS, Sullivan PF, et al. A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. Bioinformatics. 2016;32(5):650–6. https://doi.org/10.1093/bioinformatics/btv650.
111. Xu Z, Zhang G, Wu C, Li Y, Hu M. FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. Bioinformatics. 2016;32(17):2692–5. https://doi.org/10.1093/bioinformatics/btw240.
112. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24(6):999–1011. https://doi.org/10.1101/gr.160374.113.
113. Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins. 1990;7(1):41–51. https://doi.org/10.1002/prot.340070105.
114. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006;34(Web Server issue):W369–73. https://doi.org/10.1093/nar/gkl198.
115. Moses AM, Chiang DY, Eisen MB. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. Pac Symp Biocomput. 2004:324–35. https://doi.org/10.1142/9789812704856_0031.
116. Prakash A, Blanchette M, Sinha S, Tompa M. Motif discovery in heterogeneous sequence data. Pac Symp Biocomput. 2004:348–59. https://doi.org/10.1142/9789812704856_0033.
117. Sinha S, Blanchette M, Tompa M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinform. 2004;5:170. https://doi.org/10.1186/1471-2105-5-170.
118. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8. https://doi.org/10.1038/nbt.3300.
119. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011;27(12):1696–7.
120. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics. 2006;22(14):e141–e9.
121. Quang D, Xie X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–7. https://doi.org/10.1016/j.ymeth.2019.03.020.
122. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4. https://doi.org/10.1038/nmeth.3547.

123. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294–6. https://doi.org/10.1038/nmeth.2832.
124. Wang M, Tai C, Weinan E, Wei L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic Acids Res. 2018;46(11):e69. https://doi.org/10.1093/nar/gky215.
125. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–1902.e21.
126. Adey AC. Integration of single-cell genomics datasets. Cell. 2019;177(7):1677–9.
127. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell. 2019;177:1873–1887.e17.
128. Li G, Yang Y, Van Buren E, Li Y. Dropout imputation and batch effect correction for single-cell RNA sequencing data. J Bio-X Res. 2019;2(4):169–77.
129. Bengio Y. Learning deep architectures for AI. Foundations and trends® in. Mach Learn. 2009;2(1):1–127.
130. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. Adv Neural Inform Proc Syst. 2015:649–57.
131. Deng Y, Bao F, Dai Q, Wu LF, Altschuler SJ. Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. Nat Methods. 2019;16(4):311–4.
132. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8.
133. Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–729.e27.
134. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell RNA-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):1–14.
135. Way GP, Greene CS. Bayesian deep learning for single-cell analysis. Nat Methods. 2018;15(12):1009–10.
136. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Adv Neural Inform Process Syst. 2014;3:2672–80.
137. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
138. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
139. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381.
140. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740.
141. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):278.
142. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):1–12.
143. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5:2122.
144. Chen W-P, Chang S-H, Tang C-Y, Liou M-L, Tsai S-JJ, Lin Y-L. Composition analysis and feature selection of the oral microbiota associated with periodontal disease. Biomed Res Int. 2018
145. Nakano Y, Suzuki N, Kuwata F. Predicting oral malodour based on the microbiota in saliva samples using a deep learning approach. BMC Oral Health. 2018;18(1):128.
146. Hsieh C-H, Chen W-M, Hsieh Y-S, Fan Y-C, Yang PE, Kang S-T, et al. A novel multi-gene detection platform for the analysis of miRNA expression. Sci Rep. 2018;8(1):1–9.
147. Saxena D, Caufield PW, Li Y, Brown S, Song J, Norman R. Genetic classification of severe early childhood caries by use of subtracted DNA fragments from Streptococcus mutans. J Clin Microbiol. 2008;46(9):2868–73.
148. Carnielli CM, Macedo CCS, De Rossi T, Granato DC, Rivera C, Domingues RR, et al. Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nat Commun. 2018;9(1):1–17.
149. Torres PJ, Thompson J, McLean JS, Kelley ST, Edlund A. Discovery of a novel periodontal disease-associated bacterium. Microb Ecol. 2019;77(1):267–76.
150. Vapnik V. The nature of statistical learning theory. Berlin: Springer Science & Business Media; 2000.
151. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AICHE J. 1991;37(2):233–43.
152. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep. 2020;10(1):1–9.
153. Reiman D, Metwally A, Dai Y, Sun J. PopPhyCNN: a phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. IEEE J Biomed Health Inform. 2020;24(10):2993–3001.

Machine (Deep) Learning and Finite Element Modeling

14

Yan-Ting Lee, Tai-Hsien Wu, Mei-Ling Lin, and Ching-Chang Ko

14.1 Introduction

Finite element analysis (FEA) uses a numerical technique called the finite element method (FEM) to perform a simulation of a given physical phenomenon such as static structural analysis, steady-state thermal analysis, and dynamic analysis. In the following paragraph, we will briefly introduce the process and theory of traditional FEA and highlight the challenges of its use in clinical applications.

A physical system with continuum mass involves components of *geometry* and *material properties*. The phenomena of the system can be expressed by complex mathematical models, which consist of *partial differential equations* (PDEs) and *boundary conditions (BCs)*. The PDEs and BCs form a boundary value problem (BVP), and the solution of the BVP comprehensively represents the responses of the physical system. Those PDEs are called the *governing equations* of the corresponding physical problem. For example, in structural analysis problems, the governing equations are the equilibrium of forces, Hooke's law, and strain-displacement relations. The BCs can be surface load, gravity, zero displacement, etc., and the responses are displacement, strain, and stress of the structural object.

The *analytical solution* (i.e., Green's function) to a BVP is typically only available for those structures with very simple geometries and BCs. In other words, analytical solutions are often inaccessible in practical problems in which geometries and BCs are complicated. Under this circumstance, computational simulation methods such as the FEM are required to solve the BVP numerically. The FEM divides the *geometry* into smaller elements (e.g., triangles) connecting by nodes (e.g., vertices). The discretization process is called *meshing*. Thus, the governing equations become a system of algebraic equations, and the

Y.-T. Lee
Division of Oral and Craniofacial Health Sciences, Adam School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

T.-H. Wu · C.-C. Ko (✉)
Division of Orthodontics, College of Dentistry, The Ohio State University, Columbus, OH, USA
e-mail: ko.367@osu.edu

M.-L. Lin
Department of Occupational Therapy, School of Health Professions, University of Texas Health Science Center at San Antonio, San Antonio, TX, USA

C.-C. Ko et al. (eds.), *Machine Learning in Dentistry*,
https://doi.org/10.1007/978-3-030-71881-7_14

solutions of the system of equations represent the responses on all the nodes and elements. Also, we can obtain the solution anywhere in the geometry based on interpolation (shape function) on the neighboring nodes and elements, implying that the complicated mathematical model of a system is fully solved numerically. In dentistry, researchers have utilized the FEA to perform biomechanical analyses for complex dental structures including braces, wire, enamel, dentin, restoration, pulp, periodontal ligament (PDL), and Lamina dura [1, 2].

There are two ways to generate FEA geometries. First is through computer-aided design (CAD) software. Three-dimensional dental geometries, such as archwire, dental implants, and braces, can be created in CAD software. Figure 14.1 is an example where a four-tooth birth-death geometry with brackets and an archwire with 0.45 mm step bending to intrude the lateral incisor is designed using Solidworks (https://www.solidworks.com/). Another way to prepare FEA geometries is by reconstructing the scanned patient images such as computerized tomography (CT) and magnetic resonance image (MRI). Notably, the reconstruction of these images is extremely time consuming, costly, and prone to user error, and dependent on variables such as image quality.

Moreover, material properties are measured from true experimental data or simplified based on reasonable assumptions. Using true experimental data of material properties contributes to a more realistic simulation results, but can be more computationally expensive and might even result in convergence issues.

Furthermore, meshing (or discretization) helps reduce the degree of freedom from infinite to finite. The designation of different meshes is highly dependent on user knowledge regarding FEM and element types. For example, a coarse mesh can reduce the computational loading but result in inaccurate and nonmeaningful results. In contrast, a fine mesh can generate more accurate results due to more nodes and elements but increase the computational loading. Thus, different types of meshes should be assigned for different simulation problems to manage a balance between accuracy and computing speed.

Lastly, the set-up of boundary conditions is based on the corresponding physical problem. As long as the corresponding physical problem is too complicated, researchers would simplify the boundary condition based on acceptable assumptions. For example, the zero-displacement assumption is applied to the top of a four-tooth birth-death FEA model as shown in Fig. 14.1. Although the top area might have some movement in reality, it is ignored because the displacement is too small and far away from the area of interest. Based on the Saint-Venant's principle, the exact distribution of a load is not important far away from the loaded region, as long as the resultants of the load are correct. In the same example, the boundary condition on the midsagittal plane is symmetry restraint, which reduces the computational loading.

To sum, the barriers of using FEA in dental applications include long simulation time and user-dependent experience-based preprocessing setup. To address these obstacles, researchers are proposing machine learning (ML) methods to enhance the FEA processing.

14.2 Machine Learning for Geometry in Finite Element Analysis

Patient-specific FEA studies typically require manual geometry reconstruction and mesh generation of a finite element (FE) model. This process is extremely time-consuming and recommended to be performed by field experts to minimize errors. These limitations restrict FEA from time-sensitive clinical applications and studies with a large patient sample size. To overcome this challenge, Liang et al. developed a computational modeling framework based on machine learning algorithms to automate the geometry reconstruction process [3]. In their

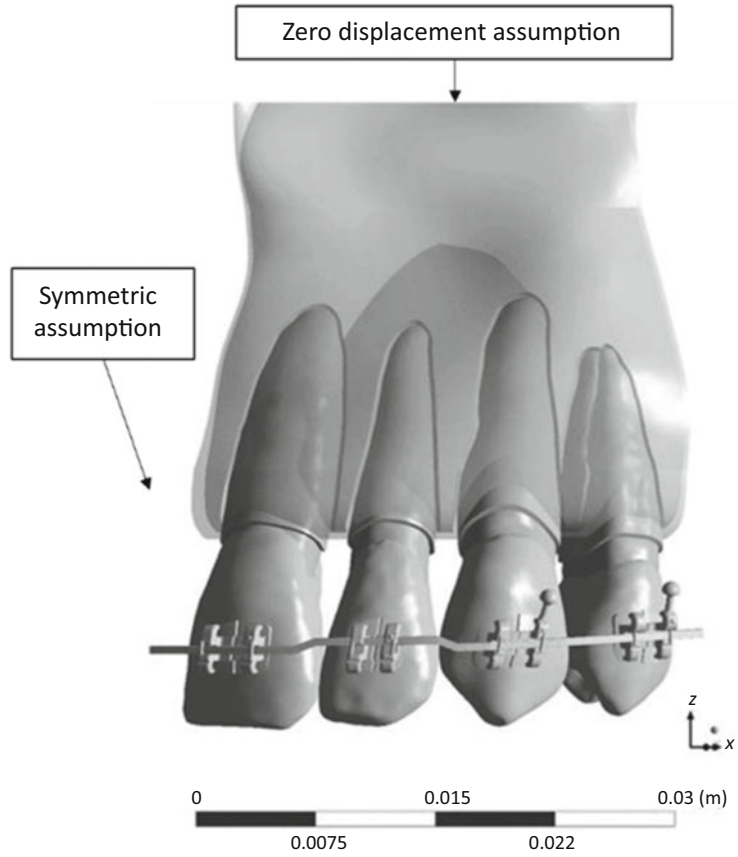


Fig. 14.1 A four-tooth birth-death FEA model with brackets and an archwire with 0.45 mm step bending to intrude the lateral incisor. Two boundary conditions are illustrated in this example. The zero-displacement assumption is used on the top of the model, and a symmetric assumption is used on the midsagittal plane. Reprinted from Ref. [2] with permission the Edward H. Angle Education and Research Foundation

study, they reconstructed the 3D geometries of the aortic valve from CT images using a machine learning based image analysis method. The reconstructed geometries can be directly used as an FE model. The mean discrepancy when comparing 10 patients' geometries reconstructed by experts using automatic reconstruction was 0.69 mm. The valve leaflet landmark and attachment curves on the root surface were detected. The FE models of valve leaflets were reconstructed by modeling fitting. They performed aortic valve closure FEA on 7 patients with the geometry generated by both manual and automatic reconstruction. The mean discrepancy for the deformed leaflet geometry in FEA was 1.57 mm. These results demonstrate a promising potential that the FE model generated by a computational framework based on machine learning algorithm can substitute the time-consuming model generation by human.

14.3 Using Machine Learning to Overcome the Computational Expenses in Finite Element Analysis

In Sect. 14.2, we discuss how research using ML to generate geometry for FEA. In this section, we review two studies that use ML to overcome a prevalent disadvantage of FEA: expensive computation. Critically thinking, can we bypass the time-consuming simulations and obtain the result directly?

The work by Liang et al. [4] answered the question. In 2017, Liang et al. studied the relationship between shape features and numerically predicted the risk of ascending aortic aneurysm (AsAA) using ML approach [4]. They reconstructed 25 AsAA patients' aorta shapes from 3D CT images. They developed a statistical shape model (SSM) based on principal component anal-

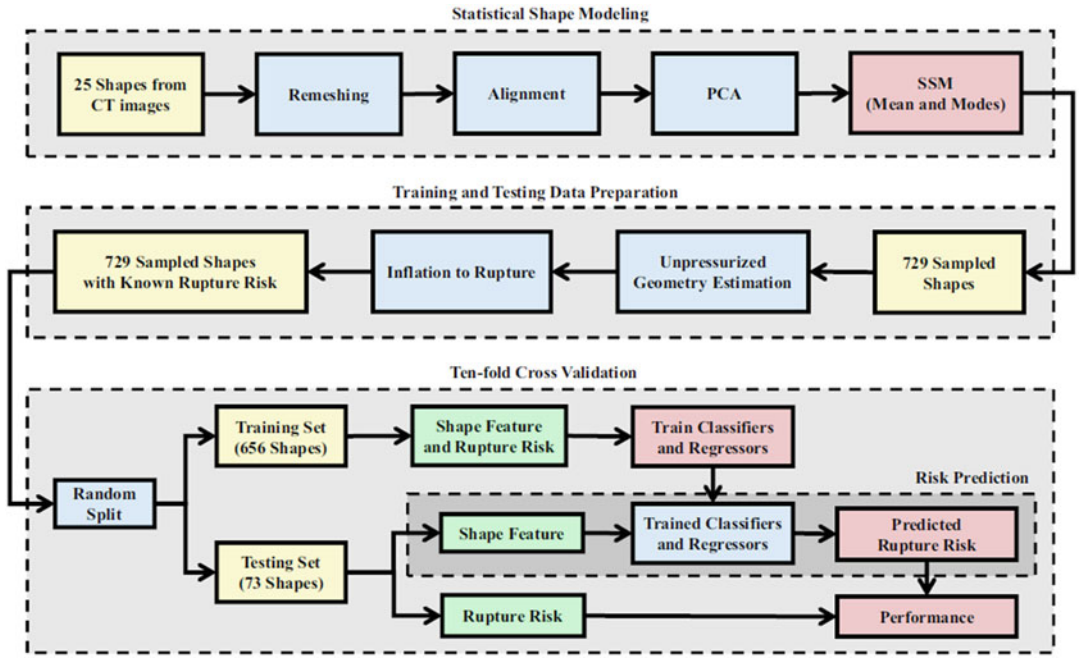


Fig. 14.2 Diagram of the modeling, simulation, and evaluation process in Ref. [4]. Reprinted from Ref. [4] with permission the Springer Nature

ysis (PCA) to describe the distribution of shapes across the study sample. Figure 14.2 shows the diagram of their work. A total of 729 representative shapes were generated with quad-surface meshing with 5000 nodes and 4950 elements. An FEA was conducted on each of the representative geometries with the material property based on their previous experiment [5]. The risk score was determined through the FEA results incorporated with the experimental study and the improved backward displacement method [6–10]. The relationship between shape feature and the risk score was determined by support vector machine (SVM) and support vector regression (SVR) [11, 12]. A ten-fold cross-validation was performed to validate the machine learning outcome. The average risk of classification accuracy was 95.58% based on an SVM classification using SSM parameters like the shape features. The average regression error was 0.0332 by using the same parameters with SVR regression. Their results show that using SSM parameters as strong shape features to make predictions of risk scores is consistent with FEA. To perform an FEA process, it usually takes about 30 minutes on average. The processing will be even longer if numerical convergence issues occurred. In contrast, SVM and SVR only take a few seconds to provide the prediction. This example shows the potential of ML approach to solving an FEA-based classification question. However, is it possible to surrogate FEA with ML approach but still obtain the FEA results (for example, the stress distribution map)?

The same research team answered the question later. In 2018, Liang et al. further developed a deep learning (DL) model to directly estimate the stress distributions of aorta [13]. In their study, they combined the previous work of ML-based geometry reconstruction and the ML-based mechanical analysis to provide a fully automatic patient-specific modeling workflow to meet the need for time-sensitive clinical applications. With the input of anatomical model, this ML-based mechanical analysis directly outputs FEA stress analysis results. They designed three DL modules which were trained separately to achieve this task. The first module, shape encoding, is a neural network that takes the shape as an input and

encodes it to a shape code. The parameter of this module is obtained by PCA. The second module, nonlinear mapping, is a multilayer neural network that maps the shape code to the stress code. The third module, stress decoding and encoding, is a bidirectional neural network with multiple layers and linear units. The stress distribution on aortic wall from the FEA can be encoded to a stress code, and the stress code from nonlinear mapping can be decoded to stress distribution in this module. The parameters of this module were obtained by low-rank approximation (LRA). The performance of the entire DL system was evaluated in ten-fold cross-validation. The average errors for Von Mises stress distribution and peak Von Mises stress were 0.492% and 0.891%, respectively. These results demonstrate that the ML approach can accelerate the computation without sacrificing the accuracy, compared to the traditional FEA results.

14.4 Using Finite Element Analysis Results as One Feature in Machine Learning for the Classification

In Sect. 14.3, we indicate the possibility of training a ML model to classify the risk of a disease to bypass the time-consuming FEA process. In this section, we further introduce another use of a combination of FEA and ML that treats FEA results as input features for a classification prediction.

In 2013, Nishiyama et al. used quantitative computed tomography (QCT) and FEA to classify women with and without hip fracture [14]. Twenty women with femoral neck fractures and 15 women with trochanteric fractures were scanned with QCT at the hip. Thirty-five age-matched controls were also scanned. The FEA was performed to estimate bone stiffness and bone failure load under a typical sideways fall on the hip. The SVM and ten-fold cross-validation were used in their classification task. When using only FEA for classifying the women with and without fracture, the accuracy was 82.9%. Combining the FEA with the volumetric bone mineral density measured at the femoral neck and total hip, the accuracy was 91.4%. Their study shows that the result of FEA itself can be an outstanding feature to serve as a classification task in machine learning. The accuracy could be further improved by the combination of FEA results with other features.

14.5 Discussion and Conclusion

In this chapter, we briefly described the theory and process of FEA and a few potential ways to combine the FEA with ML. It is important to note that although the applications reviewed in this chapter were in other medical fields but not directly in dentistry, the techniques used in other medical applications can set a foundation for the utilization of combined FEA and ML methods in dental applications. For example, Liang et al. used ML to automatically generate the aortic wall. The same technique might be applicable to automatically generate PDL which is difficult to obtain from CT or Cone beam CT for dental simulation models. In addition, using ML to surrogate FEA is another promising path, which can dramatically reduce the computational time and obtain the results in real time. With emerging digital dentistry (i.e., orthodontic aligner therapy), ML-FEA may offer opportunities to streamline the mechanical design for an individual patient. However, compared to the other medical fields using the FEA with ML, the geometries and boundary conditions in dental problems are usually more complicated, suggesting that more adjustments to the combined process are needed prior to its application in dental problems. Revisiting the question proposed in Sect. 14.3, we conclude that the combination of the FEA and ML exhibits its strength in improving the disadvantages of FEA (e.g., long computing time and user-dependent knowledge). In addition, training data is the most important role in machine learning, particularly in deep learning. The results of FEA also can be considered as a type of training data so that the combination of the FEA and ML also enhances the accuracy of prediction in classification problems.

References

1. Trivedi S. Finite element analysis: a boon to dentistry. J Oral Biol Craniofacial Res. 2014;4:200–3. https://doi.org/10.1016/j.jobcr.2014.11.008.
2. Uhlir R, Mayo V, Lin PH, et al. Biomechanical characterization of the periodontal ligament: orthodontic tooth movement. Angle Orthod. 2016;87:183–92. https://doi.org/10.2319/092615-651.1.
3. Liang L, Kong F, Martin C, et al. Machine learning–based 3-D geometry reconstruction and modeling of aortic valve deformation using 3-D computed tomography images. Int J Numer Method Biomed Eng. 2017;33:e2827. https://doi.org/10.1002/cnm.2827.
4. Liang L, Liu M, Martin C, et al. A machine learning approach to investigate the relationship between shape features and numerically predicted risk of ascending aortic aneurysm. Biomech Model Mechanobiol. 2017;16:1519–33. https://doi.org/10.1007/s10237-017-0903-9.
5. Martin C, Sun W, Pham T, Elefteriades J. Predictive biomechanical analysis of ascending aortic aneurysm rupture potential. Acta Biomater. 2013;9:9392–400. https://doi.org/10.1016/j.actbio.2013.07.044.
6. Weisbecker H, Pierce DM, Holzapfel GA. A generalized prestressing algorithm for finite element simulations of preloaded geometries with application to the aorta. Int J Numer Method Biomed Eng. 2014;30:857–72. https://doi.org/10.1002/cnm.2632.
7. Raghavan ML, Ma B, Fillinger MF. Non-invasive determination of zero-pressure geometry of arterial aneurysms. Ann Biomed Eng. 2006;34:1414–9. https://doi.org/10.1007/s10439-006-9115-7.
8. Lu J, Zhou X, Raghavan ML. Computational method of inverse elastostatics for anisotropic hyperelastic solids. Int J Numer Methods Eng. 2007;69:1239–61. https://doi.org/10.1002/nme.1807.
9. Gee MW, Förster C, Wall WA. A computational strategy for prestressing patient-specific biomechanical problems under finite deformation. Int J Numer Method Biomed Eng. 2010;26:52–72. https://doi.org/10.1002/cnm.1236.
10. Martin C, Sun W, Elefteriades J. Patient-specific finite element analysis of ascending aorta aneurysms. Am J Physiol Circ Physiol. 2015;308:H1306–16. https://doi.org/10.1152/ajpheart.00908.2014.
11. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27.
12. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/BF00994018.
13. Liang L, Liu M, Martin C, Sun W. A deep learning approach to estimate stress distribution: a fast and accurate surrogate of finite-element analysis. J R Soc Interface. 2018;15:20170844. https://doi.org/10.1098/rsif.2017.0844.
14. Nishiyama KK, Ito M, Harada A, Boyd SK. Classification of women with and without hip fracture based on quantitative computed tomography and finite element analysis. Osteoporos Int. 2014;25:619–26. https://doi.org/10.1007/s00198-013-2459-6.