Bu *et al. Progress in Orthodontics* (2025) 26:50
https://doi.org/10.1186/s40510-025-00598-8

Progress in Orthodontics

RESEARCH Open Access

Accuracy comparative study of automatic landmarking and diagnostic models on lateral cephalograms

Wen-qing Bu[1,2], Zhan-yi Shi[3], Zhi-qiang Tian[3], Meng-qi Han[1,2], Dong Zhang[3], Shuo Wang[3], Yi Sun[4], Qin Zhou[1,2], Zhi-ping Hu[1], Fei Wang[1,2], Shao-yi Du[3*] and Yu-cheng Guo[1,2,3*]

Abstract

Background The application of deep learning techniques in cephalometric analysis has become increasingly prominent. Although automatic landmarking models for cephalometric analysis have been developed, their accuracy still requires validation and relies heavily on clinicians to resolve discrepancies between results. To address these limitations, automatic diagnostic models have gained attention. However, there is no direct evidence establishing the superiority of one model over the other, especially the generalization and transferability.

Methods Based on the same northern Chinese population external test dataset data and the data of the IEEE (Institute of Electrical and Electronics Engineers) 2015 ISBI (International Symposium on Biomedical Imaging) Grand Challenge dataset, we compared the performance, generalization ability, and transfer ability of the proposed two models, respectively.

Results Our findings suggest that the automatic landmarking model outperforms the automatic diagnostic model in both external test dataset, with an accuracy of 90.80% on the IEEE dataset.

Conclusions In this study, the comparison was indirect, with each model having its strengths: the automatic landmarking model offers precise measurements, while the automatic diagnostic model provides quicker results. The choice between them depends on clinical needs, and future work should explore hybrid models to combine the advantages of both.

Keywords Deep learning, Convolutional neural networks, Artificial intelligence, Diagnostic systems, Digital imaging, Decision-making

*Correspondence:
Shao-yi Du
dushaoyi@xjtu.edu.cn
Yu-cheng Guo
xjtu-guoyucheng@163.com
[1]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, Xi'an, China
[2]Department of Orthodontics, College of Stomatology, Xi'an Jiaotong University, Xi'an, China
[3]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China
[4]Department of Imaging & Pathology, Biomedical Sciences Group, KU Leuven & Oral and Maxillofacial Surgery, University Hospitals Leuven, KU Leuven, Leuven, Belgium

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

The cephalometric analysis is regarded as the most important tool used by orthodontists and maxillofacial surgeons to study malocclusion and maxillofacial bone differences [1, 2]. Manual tracing is the primary method used in traditional cephalometric analyses. However, this method has the drawbacks of low efficiency, low precision, high time consumption, and being laborious [3].

Deep learning is a key branch of machine learning, utilizing multi-layer neural networks to extract features from complex data. It has shown great potential in medical image analysis, especially for automatic landmark detection and diagnostic tasks. With the rapid development of deep learning, artificial intelligence (AI) has gradually penetrated the medical field, which not only improves the visual recognition ability of radiology, but also greatly reduces the probability of observer error [4, 5]. Owing to the advantages of such methods, deep learning-related methods have been widely applied in medical imaging tasks, such as disease classification [6–9], ROI (region of interest) segmentation [10–12], medical detection [13–16], image registration [17–19], and achieved satisfactory results. Convolutional neural networks (CNN) are a type of deep learning algorithm that is highly effective in image processing and recognition tasks, including the automatic landmarking of lateral cephalograms [20]. Kim et al. [21] used a cascade CNN to identify lateral cephalometric landmarks across multiple centers in South Korea. Zhang et al. [22] further evaluated the performance of deep learning methods on a multi-center dataset, highlighting the challenges still faced in generalizing these methods. Deep learning models have the advantage of higher landmark accuracy, requiring lesser time and manpower when locating and analyzing anatomical markers compared to that in traditional methods. Although the use of AI technology has improved the efficiency of landmarking and measurements, a manual diagnosis using numerical values is still required. For better automation, objectivity, and accuracy in orthodontic clinical diagnosis, recent research has focused on deep learning-based end-to-end classification for skeletal malformation diagnosis, with better results [23–25]. The end-to-end model is more inclined toward self-recognition and detection of images by machines and logical association with labeled classification. The output results are not numerical values, but are diagnostic. Therefore, the end-to-end method of cephalometric measurement analysis not only solves the problem of time-consumption and high knowledge and the technical requirements of traditional methods, but also solves the problem of the automatic landmarking model, making it difficult to diagnose disputed value samples. For instance, Chang et al. [26] proposed a self-supervised learning framework for multi-attribute classification to enhance the generalization of orthodontic diagnostic models. However, there is currently a lack of direct evidence comparing the advantages of automatic landmarking models and automatic diagnostic models in the context of lateral cephalogram analysis, particularly regarding their generalization and transferability across multi-center datasets. Therefore, this study will compare the accuracy of automatic landmarking and automatic diagnostic models and explore whether AI can be directly diagnosed without requiring landmarking measurement data to clarify the development direction for AI-assisted orthodontic diagnosis.

To ensure a fair comparison, this study used two external test datasets and evaluated the performance of both models based on standardized metrics, including accuracy and generalization ability. One of them is a uniform test set sample (which comes from the same source as the model training sample) should be used for testing in this experiment to compare the performance. Meanwhile, to ensure the objectivity of comparison, we also used lateral cephalograms from the public test dataset IEEE (Institute of Electrical and Electronics Engineers) 2015 ISBI (International Symposium on Biomedical Imaging) Grand Challenge dataset [27, 28] to test the above two models. However, it is generally difficult to get enough training set samples from different populations. Thus, in the previous studies mentioned above, the samples for model construction were mostly from a single population sample, which means that the generalization ability [29] and transfer ability [30] of the model need to be studied to evaluate the applicability to other populations. Since the lateral cephalograms in the IEEE dataset were all from Han Chinese South (CHS), the generalization ability and transfer ability of the two models could also be better tested, which provides a direction for the automated diagnosis and treatment applicable to other oral diseases in the future.

Therefore, the aim of this study was to compare the accuracy, generalization ability, and transfer ability of the automatic landmarking and automatic diagnostic models using the same types of external test datasets to provide a direction for the future development of the oral diagnosis model.

Materials and methods

Datasets

Lateral cephalograms were collected from patients from northern China at the Hospital of Stomatology, Xi'an Jiaotong University, all of which were acquired using the KaVo OP 3D Pro (KaVo, Germany), a cephalometric imaging system. This device is a horizontally scanning unit, providing cephalometric images with resolution ranging from 4095×4095 pixels to 651×651 pixels. These images were manually landmarked and used for training

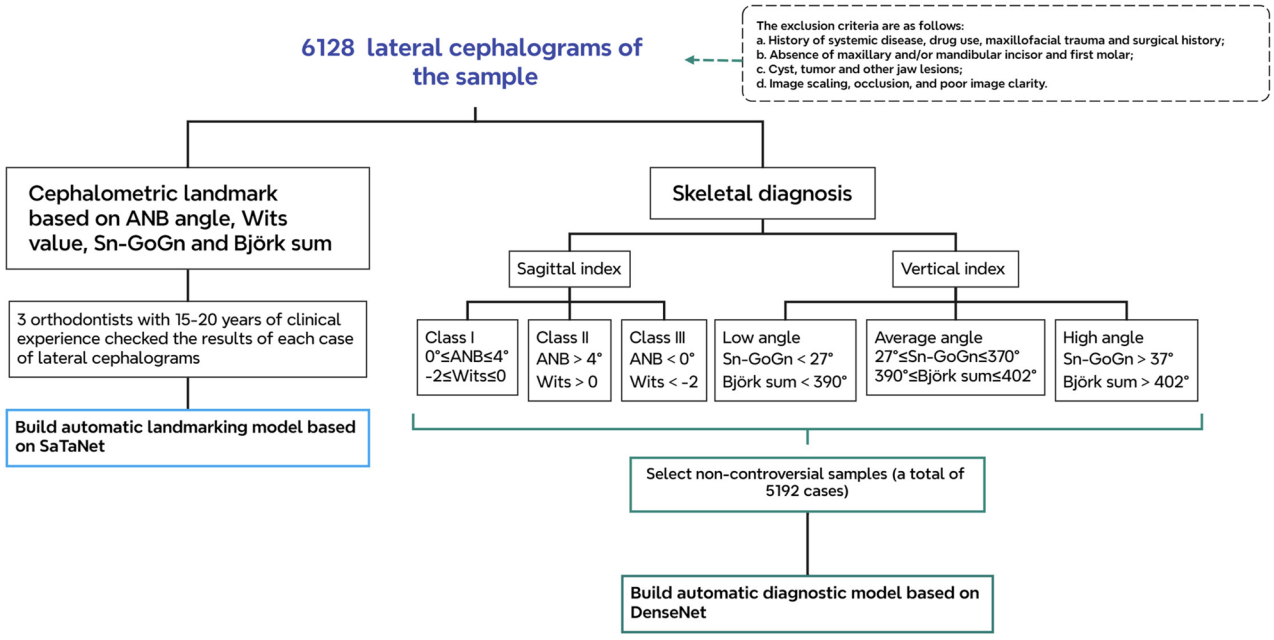


Fig. 1 The main building process of the automatic landmarking model and the automatic diagnostic model [41–43]

and testing the models (Fig. 1). The dataset was divided into two parts: the internal dataset and external test datasets (Table 1).

Internal dataset

The internal dataset was mainly used for models' establishing. The age and sex distributions of all samples are shown in Table 2.

External test datasets

Due to the random distribution of the internal dataset into different subsets, it is not suitable for a direct, final performance comparison. Therefore, an external test dataset was introduced to independently evaluate the models' performance. To evaluate the model's performance and clinical applicability, we used a separate external test dataset consisting of 426 lateral cephalograms. These images were collected from the Hospital of Stomatology, Xi'an Jiaotong University, ensuring they were independent of the internal dataset. This external dataset could offer a more realistic comparison. Additionally, to further test the generalization and transferability of the models across different populations and imaging systems, we included the IEEE 2015 ISBI Grand Challenge dataset [27, 28] as part of the external test set. The IEEE dataset includes 250 cephalograms from Taiwan Province of China and serves as a widely-used benchmark for evaluating cephalometric landmarking algorithms. This allows us to evaluate the models' ability to generalize across diverse clinical settings.

Table 3 Location and definition of landmarks on lateral cephalograms

Number	Label	Definition
1	R1	Ruler point (upper)
2	R2	Ruler point (20 mm below R1)
3	N	Nasion
4	S	Sella
5	Ar	Articulare
6	Go	Gonion
7	Po	Pogonion
8	Gn	Gnathion
9	Me	Menton
10	A	Subspinale
11	B	Supramental
12	O1-O2	Occlusal plane

Automatic landmarking model

All 6128 lateral cephalograms in the internal dataset were used to establish the model. Table 3 lists the 12 cephalometric landmarks (a total of 12) of lateral cephalograms. In this section, a CNN named SaTaNet (structure-aware and texture-aware CNN-based end-to-end network) has been used [31], which combines structure awareness and texture awareness. In the overall framework of SaTaNet (Fig. 2), the structure awareness and texture awareness integrated module (SaTa) consists of a trunk (its encoder is ResNeXt-50, pre-trained on ImageNet) and three heads based on the CNN network. The three heads are the structure awareness head (to guide the network to localize the landmarks coarsely), the CLS head (classification head, helping to confirm the presence and precise location of the landmarks by classifying pixels as either

Table 4 The definition of measurement indices

Dimension	Index	Definition
Sagittal	ANB angle	Formed by A point, nasion, and B point, which describes the anteroposterior relationship of the mandible to the maxilla
	Wits value	Measured by drawing a line perpendicular to the occlusal plane from the midpoint of the maxillary incisor to the mandibular incisor for evaluating the anteroposterior relationship between the upper and lower jaws
Vertical	Sn-GoGn	The angle between the Sella-Nasion line and the Gonion-Gnathion line for evaluating the inclination of the mandible about the cranial base
	Björk sum	The sum of the anterior and posterior facial heights

landmarks or not), and the texture awareness head (for fine localization of the landmarks). The structure and texture information of the sample guided the SaTa module for structure and texture perception, respectively.

For model training, all cephalometric images were preprocessed before being fed into the network. Each image was resized to a unified resolution with zero-padding to maintain the original aspect ratio, and pixel intensity was normalized. To improve robustness and reduce overfitting, data augmentation strategies such as random rotation, flipping, and scaling were applied. The SaTaNet model was optimized using the Adagrad optimizer with a batch size of 16. The loss function consisted of three parts: a structure-aware loss (L_{str}), a texture-aware loss (L_{tex}), and a pixel-wise classification loss (L_{cls}), enabling both coarse localization and fine-grained refinement of landmarks [31]:

The total loss function (L_{total}) can be defined as follows:

$$L_{total}\left(LM,\ \mathbb{C}\left(\theta\ ,I\right)\right)=\alpha\ L_{str}+\beta\ L_{tex}+L_{cls} \quad (1)$$

the α and β are the balancing factors (set α and β to 100). The θ is the parameter of the network $\mathbb{C}$, and the LM is the annotated landmarks in image I.

Texture-aware loss can be defined as follows:

$$L_{tex}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{X}\sum_{k=1}^{Y}|H_{tex}\left(j,\ k\right)-\widehat{H}_{tex}\left(j,\ k\right)|}{N\times\ X\times\ Y} \quad (2)$$

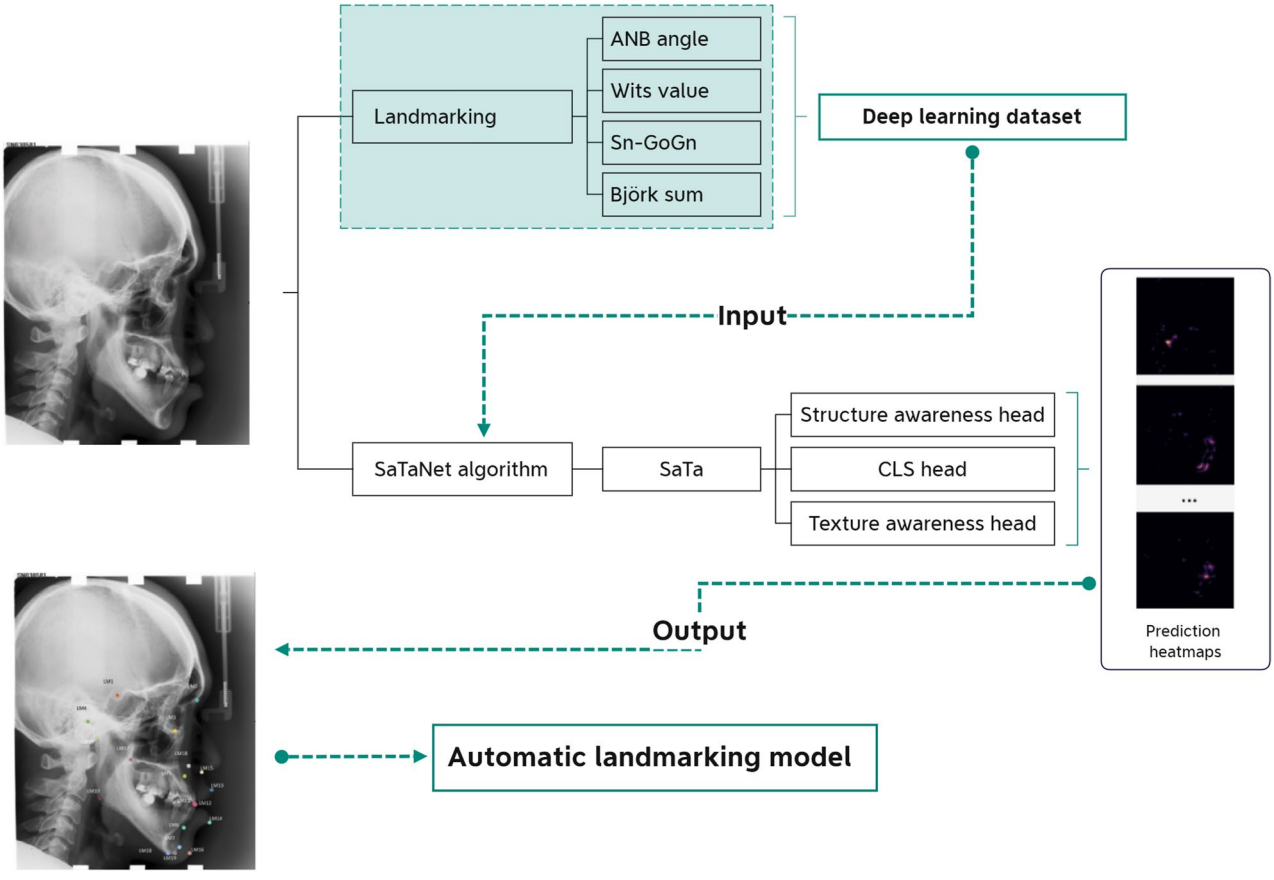


Fig. 2 The general framework of the automatic landmarking model based on the SaTaNet algorithm [31]

where N is the number of landmarks, X and Y are the image width and height, H_{tex} is the corresponding texture label produced by texture label generation algorithm for an image I, $\widehat{H}_{tex}$ is the predicted texture heatmap, and i is the landmark index.

Structure-aware loss can be defined as follows:

$$L_{str} = \frac{\sum_{j=1}^{X} \sum_{k=1}^{Y} |H_{str}(j,\ k) - \widehat{H}_{str}(j,\ k)|}{X \times Y} \quad (3)$$

where H_{str} is the corresponding structure label produced by texture label generation algorithm for an image I, and $\widehat{H}_{str}$ is the predicted structure heatmap.

Pixel-wise classification loss can be defined as follows:

$$L_{cls} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{X} \sum_{k=1}^{Y} \left[-H_{cls}(j,\ k) \times \log\left(\hat{H}_{cls}(j,\ k) \right) \right] \quad (4)$$

where H_{cls} and $\widehat{H}_{cls}$ are the corresponding pixel-wise classification label for an image I and predicted classification heatmap.

The heat map prediction part of the model uses the above three heads for the visual output of the calculation results and extracts the hottest point to obtain the predicted landmarks of the hard-tissue anatomical structure of the sample (Fig. 3). According to the prediction coordinates, the ANB angle, Wits value, Sn-GoGn, and Björk sum were automatically output (Table 4). Also, we evaluated the model's performance by using Mean Radial Error (MRE) and Success Detection Rate (SDR).

Automatic diagnostic model

The samples used were all non-controversial samples (cases where expert orthodontists provided differing classifications, or cases where the measured indices fall near the classification boundaries) from the internal dataset with clinical real diagnosis as to eliminate the influence of subjective judgment as much as possible for the automatic diagnostic model. In this dataset, the distribution of Class I, Class II, and Class III was 2149, 2638, and 405, respectively. The distribution of low-angle, average-angle, and high-angle samples was 588, 2897, and 1707, respectively. This method is a medical image classification approach based on the DenseNet network, aimed at improving the classification accuracy and diagnostic effectiveness of medical images. First, during the data preparation and processing phase, lateral cephalograms were collected, measured and resized to ensure uniformity, using real diagnosis results for preprocessing and pattern matching to reduce the impact of irrelevant factors in the data. Then, a batch size of 32 was used, and the Adam optimizer was employed for parameter optimization. To enhance the diversity of the dataset, a series of image transformation strategies were applied, including scaling adjustment, brightness enhancement, center rotation, random occlusion, edge detection, and mirror transformation. And the processed data was input into the Dense Block feature extraction module, which realized feature reuse through the connection of features on the channel by establishing dense connections between all the previous layers and the following layers. Subsequently, the nested supervision module was incorporated after the first layer based on the Dense Block, and the channel attention module (Squeeze-and-Excitation Block, SEBlock) was appended after the third layer. The soft threshold automatically learned by the nested supervision module was employed to allocate the weight of each channel more appropriately, enabling the module to automatically learn and apply the threshold prior to the fully connected layer to achieve the soft thresholding of the feature map. This facilitates the network to process image features with greater accuracy. Next, the fourth layer of the Dense network (Dense Block4 and Transition Layer4), was connected to achieve the prediction classification results. Moreover, cross-entropy loss

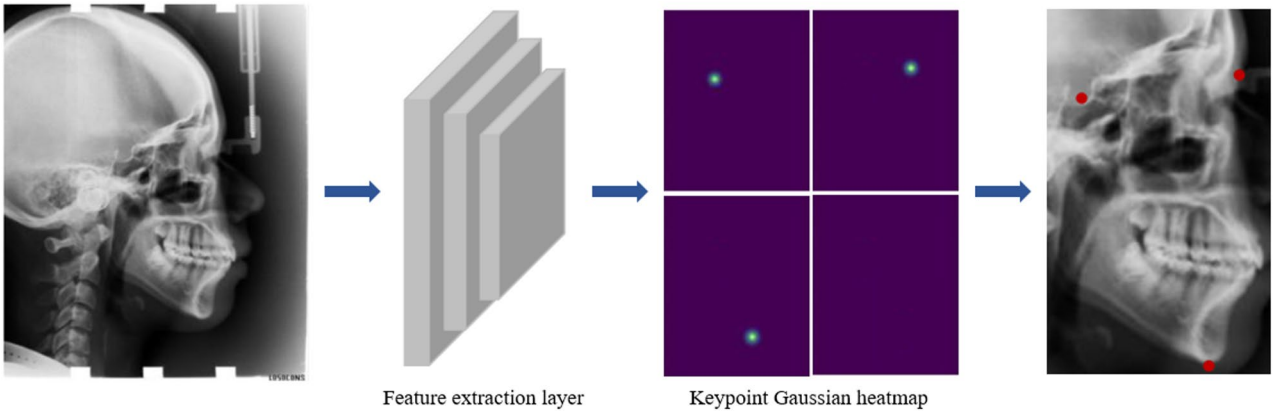


Fig. 3 Image keypoint detection method based on heatmap prediction (taking points S, N, and Me as examples)

was employed in model training to optimize the model parameters:

$$Loss = -\beta \log \frac{exp\,(z_j)}{\sum_{i=1}^{c} exp\,(z_i)} \quad (5)$$

where z_j and z_i represent the output of the network, β is a re-weighting factor for the loss function, c is the number of classes.

By integrating the advanced channel attention mechanism and the deep network features of DenseNet, this approach effectively enhanced the classification accuracy and robustness of medical images in the presence of noise, thereby improving the diagnostic efficacy. Finally, according to the input images, the sagittal (Class I, Class II, or Class III) and vertical (low-angle, average-angle, or high-angle) predicting diagnosis results were output. (Fig. 4) In addition, to further investigate the model's decision-making in the classification task, we used Grad-CAM to visualize the model's attention regions, highlighting key areas of feature extraction and classification.

Comparison of the two models

Total of 426 lateral cephalograms of the northern Chinese population were from the external test dataset. The images were manually landmarked, diagnosed, and verified respectively. This dataset includes samples with controversial values that can be used to test the clinical applicability of the model. To facilitate the comparison between the two networks, we classified the automatic landmarking model's predicted measurements into clinically relevant categories based on their corresponding ANB, Wits, Sn-GoGn, and Björk values. The automatic diagnostic model output directly provides sagittal and vertical diagnoses, which are compared to the automatic diagnostic model's categorized outputs. For evaluation, we used standard performance metrics such as F1-score, accuracy, precision, and recall to evaluate the agreement between the two models. These metrics allow us to quantify the classification accuracy of each network based on the clinical categories. While these indirect measures help evaluate the general performance of both networks, we acknowledge the limitation that the automatic landmarking model's output involves multiple intermediate steps (landmark identification and measurement calculation), whereas the automatic diagnostic model outputs the final diagnostic categories directly. This difference requires careful consideration in our discussion, especially when analyzing the two models' direct comparison.

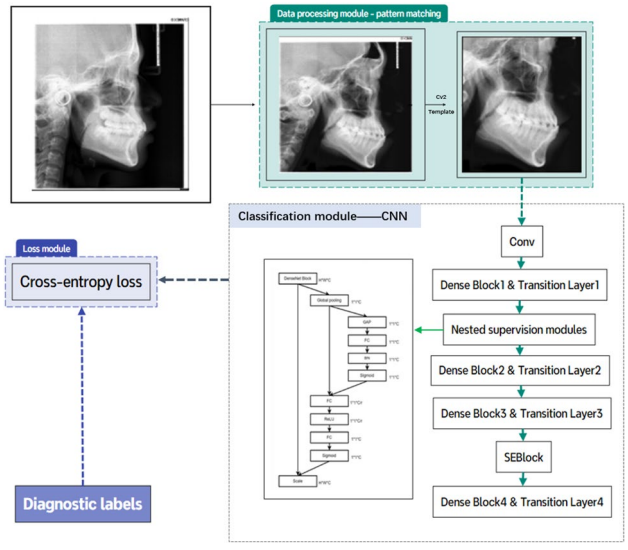


Fig. 4 The general framework of the automatic diagnostic model based on the DenseNet algorithm

The IEEE dataset is including anatomical landmarking. However, the Wits value could not be measured at the sites labeled in this dataset [27, 28]. All the included samples were simultaneously checked by three orthodontists with 15–20 years of clinical experience at the same time to determine the sagittal and vertical diagnostic classification of each sample. These results were subsequently used as the manual results of the samples to test the accuracy, precision, recall, and F1 Score of the trained models. In addition, McNemar's test was performed to evaluate whether the difference in classification performance between the two models was statistically significant on paired prediction results.

Results

The landmark detection performance of the automatic landmarking model is evaluated by MRE and SDR. In the SaTaNet model [31], the performance on the Chinese population test set showed an MRE of 1.15 mm, with an SDR of 86.80% within a 2 mm threshold. On the IEEE test set, the MRE reached 1.27 mm, and the SDR within 2 mm was 82.01%. The performance of the automatic diagnostic model is shown by the Grad-CAM results in Fig. 5. In the successfully classified cases, the model mainly focused on the key areas of the maxilla and mandible. The heatmap shows high responses (red and yellow areas) in these regions, with a relatively concentrated distribution. In contrast, in the misclassified examples, the heatmap distribution was more scattered, and in some images, the areas of focus were not correctly centered on the key maxillofacial regions, leading to incorrect classifications.

The results of the northern Chinese population external test dataset for the two models are presented in Table 5. For the sagittal index, the accuracy of the automatic diagnostic model was 84.27%, which was better than that of the automatic landmarking model and equal to that of the ANB angle. While the automatic landmarking model achieves high classification accuracy based on ANB angle, the diagnostic model offers a more efficient

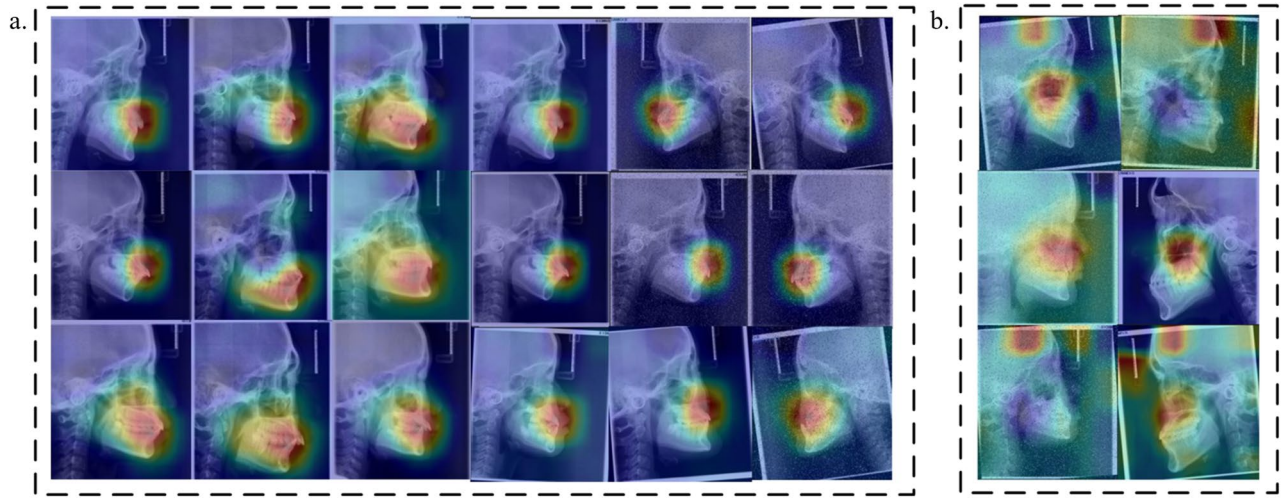


Fig. 5 Grad-CAM Visualization results based on the automatic diagnostic model: **a** Successful cases, **b** Error cases

Table 1. The functional grouping of the datasets

Datasets	Internal Dataset		External Test Datasets	
	Automatic landmarking model	Automatic diagnostic model*	The northern Chinese population	The IEEE
Training Set (n)	4044	4159	0	0
Test Set (n)	2084**	1033**	426	250
Total (n)	6128	5192	426	250

*The samples used were non-controversial samples selected from all 6128 samples

**Due to the random distribution of the internal dataset into different subsets, it is not suitable for comparison

workflow by eliminating landmarking dependency. However, its performance drops when tested on the CHS, indicating potential challenges in model generalization. For the vertical index, the test result of the automatic diagnostic model was 72.77%, which is similar to the classification performance of the Sn-GoGn and Björk sum.

The precision, recall, and F1 Score of the Wits value indicated that it was significantly less effective in the classification of Class I. Neither precision nor recall comprehensively reflected the predictive performance of the machine. The F1 Score is an index obtained by precision and recall and can comprehensively reflect the predictive performance of the machine. Therefore, the F1 Score results were mainly referred to in Ref. For the sagittal index, the F1 Score of the automatic diagnostic model was slightly lower than that of the ANB Angle. However, it was basically unchanged and significantly better than the Wits value. For the vertical index, the high angle of the automatic diagnostic model showed better F1 Score. Also, in Table 5, McNemar's test indicated that in the northern Chinese population test set, a statistically significant difference between the two models was observed only in the sagittal diagnosis based on Wits appraisal ($p < 0.05$).

The results of the IEEE test set for the two models are listed in Table 5. In the automatic landmarking model, the results of Classes II and III were better than those of Class I for ANB Angle classification. In the Sn-GoGn classification, the classification performances of the high, average, and low angles were basically the same. The Björk sum showed the best accuracy, and the classification performances of the high, average, and low angles were essentially the same. Compared with the test results of the northern Chinese population, the performance of the automatic landmarking model was still good. However, in the test results of the automatic diagnostic model, the sagittal diagnosis was slightly better than the vertical diagnosis. However, both were lower than the results

Table 2 Age and sex distribution of all samples

Age(y)	Automatic landmarking model			Automatic diagnostic model		
	Male (n)	Female (n)	Total (n)	Male (n)	Female (n)	Total (n)
8–17	1495	2106	3601	1228	1754	2982
18–27	529	1340	1869	515	1143	1658
28–37	129	426	555	112	352	464
38–47	20	69	89	14	61	75
48–57	3	9	12	3	8	11
58–67	1	1	2	1	1	2
Total(n)	2177	3951	6128	1873	3319	5192

Table 5 The results of the two external test datasets in the two models

Test set	Automatic diagnostic model						Automatic landmarking model						McNemar's Test (with each two indices)	
	Index	Diagnosis	Accuracy	Precision	Recall	F1 value	Index	Diagnosis	Accuracy	Precision	Recall	F1 value	Valid Cases (n)	Exact Sig. (2-sided)
The northern Chinese population test set	Sagittal	Class I	84.27%	65.48%	87.76%	75.00%	ANB angle	Class I	88.50%	84.02%	95.30%	89.31%	426	0.069
								Class II		95.74%	88.24%	91.84%		
		Class II		90.28%	80.25%	84.97%		Class III		99.14%	92.74%	95.83%		
							Wits value	Class I	56.57%	6.52%	6.82%	6.67%	426	0.000*
		Class III		96.15%	73.53%	83.33%		Class II		92.16%	63.51%	75.20%		
								Class III		84.26%	98.47%	90.81%		
	Vertical	Low angle	72.77%	91.43%	50.79%	65.31%	Sn-GoGn	Low angle	74.18%	97.67%	59.57%	74.01%	426	0.695
								Average angle		70.41%	76.24%	74.01%		
		Average angle		58.82%	84.97%	69.52%		High angle		70.83%	98.08%	82.26%		
							Björk sum	Low angle	77.46%	88.37%	60.32%	71.70%	426	0.125
		High angle		85.93%	78.91%	82.27%		Average angle		68.59%	59.78%	63.88%		
								High angle		47.11%	98.28%	63.69%		
The IEEE test set	Sagittal	Class I	69.60%	53.62%	94.87%	68.52%	ANB angle	Class I	83.60%	66.30%	85.92%	74.85%	250	0.000*
		Class II		88.37%	67.86%	76.77%		Class II		94.44%	76.58%	84.58%		
		Class III		92.31%	40.00%	55.81%		Class III		92.65%	92.65%	92.65%		
	Vertical	Low angle	67.60%	100.00%	16.67%	28.57%	Sn-GoGn	Low angle	88.00%	80.00%	100.00%	88.89%	250	0.000*
								Average angle		88.96%	91.33%	88.89%		
		Average angle		69.75%	65.35%	67.48%		High angle		91.07%	75.00%	82.26%		
							Björk sum	Low angle	90.80%	78.57%	97.06%	86.84%	250	0.000*
		High angle		64.29%	87.10%	73.97%		Average angle		93.41%	92.86%	93.13%		
								High angle		92.68%	79.17%	85.39%		

* McNemar's test indicates a significant difference ($p < 0.05$)

of the automatic landmarking model in the CHS. The decline in accuracy for the diagnostic model on the IEEE dataset suggests that craniofacial morphological differences between populations affect the model's ability to generalize. This highlights the need for diverse training data to improve transferability. Besides, McNemar's test showed that significant differences were found across all evaluated indices (all $p < 0.05$) in the IEEE dataset.

Discussion

Accurate diagnosis and treatment plans are the keys to successful orthodontic treatment. Lateral cephalogram measurements are a standard part of clinical records for orthodontic diagnosis and treatment planning. CNN makes the automatic landmarking mode widely studied and applied in lateral cephalogram measurements and has achieved good results [3, 32–38]. However, this kind of landmarking model used the manually marked results as the "gold standard" in deep learning, and the learned samples themselves have certain errors. In addition, controversial measurement indicators still rely on the subjective judgment of an experienced orthodontist. Therefore, the performance of a cephalometric analysis is susceptible to human factors. Therefore, the application of the end-to-end diagnostic classification model has been focused on [23–25]. Using the DenseNet al.gorithm, Li et al. [24] used 2432 lateral cephalograms with sagittal diagnoses to train an end-to-end sagittal diagnosis model with an accuracy of 89.58%. However, there is no direct evidence of the advantages of the two automatic orthodontic diagnostic models based on lateral cephalograms.

To explore the performance of the two models, the same external test dataset was used to test the automatic landmarking model based on the SaTaNet algorithm and the automatic diagnostic model based on the DenseNet algorithm. First, we used a northern Chinese population external test dataset from the same source as the two models' internal dataset. This test set sample has a critical value problem and a contradictory value problem, similar to those in clinical practice. The critical value problem occurs when the output result is between two adjacent categories, and the craniomaxillofacial bone problem is difficult to distinguish with the naked eye. A contradictory value refers to the output values of the same type of diagnostic indicator corresponding to the contradictory diagnosis. For example, in patients with mild Class II or Class III, a Class I result may be obtained during the measurement. In patients with mildly high or low angles, it is possible to obtain an average angle result in the measurement. Therefore, the external test dataset can reflect the clinical applicability while testing the efficacy of the model.

According to the results of the automatic landmarking model (Table 5), the classification efficiency and accuracy of Wits value was significantly lower than those of the other indicators, whereas the ANB angle perspective showed a better classification efficiency, which may be related to the site-point difficulty of the relevant sites of the indicators. In determining the ANB Angle, points A, B, and N have clear anatomical characteristics that are easy to identify and learn using a deep learning network. Except for points A and B, the determination of the Wits value was related to the geometric positioning of the occlusal plane. Inclination of the occlusal plane can cause significant changes in the vertical distance between points A and B, resulting in measurement errors and reduced classification efficiency. For the Sn-GoGn and Björk sum, the overlapping of images (e.g., point Ar, point Go) or regional imaginary points (e.g., point S) exist at the relevant sites, resulting in errors in the landmarking measurement and reduced efficiency. Previous studies have pointed out that, compared with the regions located at clear boundaries, the fuzzy regions of craniomaxillofacial structures are more difficult to identify and detect [39]. However, different diagnostic classifications also show different differences in the accuracy, precision, recall, and F1 Score. The closer to the normal range, the lower the classification efficiency (Table 5), which may be related to the smaller feature difference in the normal range, smaller range of average angle, and Class I in the index classification, and sample imbalance.

For the automatic diagnostic model (Table 5), the accuracy of the sagittal diagnosis classification was higher than that of the vertical diagnosis classification. Simultaneously, the classification performance of Class II and high-angle samples was better, which was related to the unbalanced sample size, obvious degree of skeletal malformations, and difficulty in vertical diagnosis. Kim et al. [40] collected 1574 lateral cephalograms of individuals aged 5.9–64 years, diagnosed the samples only by simple ANB angle, and then used the classification results as sample labels to train a craniomaxillofacial sagittal bone automatic diagnostic model. The results showed that the classification AI model based on deep learning is superior to automatic landmarking AI software when using cephalometric images for sagittal classification. This is contrary to the results of ours. The accuracy of the automatic landmarking model is intrinsically linked to the precision of cephalometric measurements. In the Grad-CAM visualization (Fig. 5), misclassified examples showed scattered heatmap distributions, with some areas incorrectly focusing on the cranial bones, midface, pharynx, cervical vertebrae, etc., leading to errors. These results show that the model typically learns skeletal features well, but struggles with attention and boundary recognition in misclassified cases. Therefore, while the automatic diagnostic model

simplifies the process by passing the measurement step, it still requires accurate landmarking to ensure reliable diagnostic results.

To explore a more objective performance comparison and evaluate the generalization and transfer abilities of the different populations of the two models trained using a single-source internal dataset, the IEEE 2015 ISBI Grand Challenge dataset [27, 28] was adopted for testing. The results show that the SaTaNet automatic landmarking model performs well in both the northern Chinese population and the IEEE external test datasets. Therefore, the automatic landmarking model based on the SaTaNet al.gorithm has good generalization and transfer ability for different populations (in F1-score, accuracy, MRE or SDR). It can achieve relatively stable and accurate landmarking and measurement efficiency in the cephalometric images of different populations, making it convenient for auxiliary diagnosis, program formulation, and evaluation in orthodontic clinics. Although there are differences in the maxillofacial morphology of different populations, the automatic landmarking model has little effect on performance. This may be because the landmarking studies focused on the detection of craniomaxillofacial anatomical sites rather than on the overall morphology, and the morphological differences had little effect on the detection efficiency of the landmarking model. Although the DenseNet automatic diagnostic model achieved good performance test results in the northern Chinese population external test dataset and solved the problem of contradictory value diagnosis, it was not satisfactory in the test of the IEEE dataset based on the CHS. Therefore, the DenseNet automatic diagnostic model proposed in the present study is more suitable for the northern Chinese population, which has a relatively poor generalization and transfer ability to other populations. One possible reason for the decreased performance of the diagnostic model in the IEEE dataset is the difference in craniofacial structures across populations. Since the model was trained predominantly on a northern Chinese dataset, it may have learned population-specific features, making it less adaptable to other ethnic groups. In this end-to-end model, the machine independently learns the mapping relationship between craniomaxillofacial morphology and diagnosis results during the training process. Therefore, the craniomaxillofacial morphology differences of different populations will affect the performance of the diagnostic model. These findings were further supported by McNemar's test. In the northern Chinese population, the two models only showed a significant difference in the sagittal diagnosis based on Wits appraisal, whereas in the IEEE dataset, the two models consistently differed across all indices. This suggests that the relative advantage of one model over the other may be dataset-dependent, highlighting the importance of validating diagnostic models across diverse populations and measurement indices.

In this study, we compared the performance of two deep learning models: the automatic landmarking model and the automatic diagnostic model. The automatic landmarking model first predicts landmark coordinates from lateral cephalograms, which are then used to calculate key cephalometric measurements, such as the ANB angle, Wits value, Sn-GoGn angle, and Björk sum. These measurements are subsequently classified into clinically relevant categories to evaluate the model's diagnostic performance. In contrast, the automatic diagnostic model directly outputs diagnostic classifications (e.g., Class I, II, III or high, average, low angles), eliminating the need for intermediate measurement calculations. Although both models are evaluated using standard performance metrics (accuracy, precision, recall, and F1-score), it is important to note the differences in their processing: the automatic landmarking model involves multiple steps—first predicting landmarks, then calculating measurements, and finally classifying these measurements—while the automatic diagnostic model directly outputs the final diagnostic classification. This difference in processing introduces potential variability in the automatic landmarking model, as the accuracy of landmark detection and subsequent measurement calculation can influence the final diagnosis. Conversely, the automatic diagnostic model's direct output avoids this variability but may provide less detailed insight into the specific measurements used for classification. We acknowledge these methodological differences and their implications for performance evaluation. While the automatic landmarking model provides a more detailed analysis of cephalometric landmarks and measurements, the automatic diagnostic model is more efficient in clinical applications, as it directly produces diagnostic categories without the need for further processing. However, this comparison must be interpreted cautiously, as the two models fundamentally differ in approach. Future work could explore a hybrid model that combines the strengths of both: integrating the landmark-based approach of the automatic landmarking model with the direct diagnostic output of the automatic diagnostic model. Such a hybrid model could offer a comprehensive solution, balancing the accuracy of landmark-based measurements with the efficiency of direct diagnostic classification.

Another limitation of this study is that the dataset for Class I, Class II, and Class III malocclusions is imbalanced, with fewer samples for Class III malocclusion, which may affect the model's predictive accuracy for Class III cases. In future studies, more cases of Class III malocclusion should be studied. However, owing to the lack of Class III malocclusion cases in the northern Chinese population, this would be a difficult and lengthy

process. Additionally, to enhance the generalization capability of both models, future work should incorporate more diverse training datasets, including samples from various ethnic backgrounds.

Conclusion

In this study, we compared the automatic landmarking model and automatic diagnostic model for automated orthodontic diagnosis. The automatic landmarking model provides detailed landmark-based measurements, while the automatic diagnostic model directly outputs diagnostic classifications. Given the differences in their outputs, the comparison is indirect, and any conclusion on accuracy depends on the specific evaluation framework used. Both models have strengths: the automatic landmarking model offers precise measurements, useful for detailed analysis, while the automatic diagnostic model is more efficient, providing rapid diagnostic results. The choice between them depends on clinical needs, and future work should explore hybrid models combining the advantages of both approaches.

Abbreviations

IEEE	Institute of electrical and electronics engineers
ISBI	International symposium on biomedical imaging
AI	Artificial intelligence
CNN	Convolutional neural networks
CHS	Han Chinese South
ROI	Region of interest
SaTaNet	Structure-aware and texture-aware CNN-based end-to-end network
SaTa	The structure awareness and texture awareness integrated module
CLS	Classification head
MRE	Mean radial error
SDR	Success detection rate
SEBlock	Squeeze-and-excitation block

Acknowledgements

Not applicable.

Author contributions

WQB, ZYS, MQH, DZ, and SW performed data analysis, investigation, and wrote the original draft. ZQT, YS, QZ, ZPH, FW, SYD, and YCG contributed to critical revision, funding acquisition, or final approval of the manuscript. SYD, YCG, FW, SW, and YS conceived and designed the study. WQB, ZYS, ZQT, and SW developed the methodology and provided resources. WQB and ZQT curated and validated the data. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 81701869, 81901926), the Shaanxi Natural Science Funds for Distinguished Young Scholars of China (Grant No. 2025JC-JCQN-101), the Key Research and Development Program of Shaanxi (Grant No. 2021SF-047), and the Fundamental Research Funds for the Central Universities (Grant No. xzy022022065).

Data availability

The authors do not have permission to share data.

Declarations

Ethics approval and consent to participate

The ethical standards of the institutional and/or national research committee and the 1964 Declaration of Helsinki and its later amendments, or comparable ethical standards, were followed throughout the study. The current research was conducted after the approval of the local Biomedical Ethics Committee (No: 2023-XJKQIEC-043-002).

Consent for publication

The patient (or legal guardian(s) of minors) provided written informed consent for the publication of anonymized clinical details and associated images in this study.

Competing interests

The authors declare no competing interests.

Received: 30 May 2025 / Accepted: 19 November 2025
Published online: 27 November 2025

References

1. Broadbent H. New X-ray technique and its application to orthodontics. Angle Orthod. 1931;1:45–66.
2. Hofrath H. Die bedeutung der Röntgenfern- und abstandsaufnahme für die diagnostik der Kieferanomalien. J Orofac Orthop. 1931;1:232–58. https://doi.org/10.1007/BF02002578.
3. Yao J, Zeng W, He T, Zhou S, Zhang Y, Guo J, et al. Automatic localization of cephalometric landmarks based on convolutional neural network. Am J Orthod Dentofacial Orthop. 2022;161:e250-9. https://doi.org/10.1016/j.ajodo.2021.09.012.
4. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, Tridandapani S, Auffermann WF. Deep learning in radiology. Acad Radiol. 2018;25:1472–80. https://doi.org/10.1016/j.acra.2018.02.018.
5. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, Flanders AE, Lungren MP, Mendelson DS, Rudie JD, Wang G, Kandarpa K. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The academy workshop. Radiology. 2019;291:781–91. https://doi.org/10.1148/radiol.2019190613.
6. Li X, Jia M, Islam MT, Yu L, Xing L. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Trans Med Imaging. 2020;39:4023–33. https://doi.org/10.1109/TMI.2020.3008871.
7. Shorfuzzaman M, Hossain MS. MetaCOVID: a siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. Pattern Recognit. 2021;113:107700. https://doi.org/10.1016/j.patcog.2020.107700.
8. Zhang Y, Jiang H, Miura Y, Manning C, Langlotz C. Contrastive learning of medical visual representations from paired images and text. ArXiv. 2020. https://doi.org/10.48550/arXiv.2010.00747.
9. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing. 2018. https://doi.org/10.1016/j.neucom.2018.09.013.
10. Alom MZ, Hasan M, Yakopcic C, Taha T, Asari V. Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. ArXiv. 2018. https://api.semanticscholar.org/CorpusID:3573942.
11. Yu L, Wang S, Li X, Fu C-W, Heng P-A. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. Cham: Springer International Publishing; 2019. p. 605–13. https://doi.org/10.48550/arXiv.1907.07034.
12. Fan DP, Zhou T, Ji GP, Zhou Y, Chen G, Fu H, Shen J, Shao L. Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans Med Imaging. 2020;39:2626–37. https://doi.org/10.1109/tmi.2020.2996645.
13. Rijthoven Mv, Swiderska-Chadaj Z, Seeliger K, Laak Jvd, Ciompi F. You Only Look on Lymphocytes Once. 2018. https://api
14. Mei J, Cheng MM, Xu G, Wan LR, Zhang H. SANet: a slice-aware network for pulmonary nodule detection. IEEE Trans Pattern Anal Mach Intell. 2022;44:4374–87. https://doi.org/10.1109/tpami.2021.3065086.
15. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. Med Image Anal. 2020;59:101557. https://doi.org/10.1016/j.media.2019.101557.

16. Zheng Y, Liu D, Georgescu B, Nguyen H, Comaniciu D. 3D deep learning for efficient and robust landmark detection in volumetric data. Cham: Springer International Publishing; 2015. p. 565–72. https://doi.org/10.1007/978-3-319-24553-9_69.
17. Simonovsky M, Gutiérrez-Becker B, Mateus D, Navab N, Komodakis N. A deep metric for multimodal registration. Cham: Springer International Publishing; 2016. p. 10–8. https://doi.org/10.48550/arXiv.1609.05396.
18. Sokooti H, de Vos B, Berendsen F, Lelieveldt BPF, Išgum I, Staring M. Nonrigid image registration using Multi-scale 3D convolutional neural networks. Cham: Springer International Publishing; 2017. p. 232–9. https://doi.org/10.1007/978-3-319-66182-7_27.
19. Balakrishnan G, Zhao A, Sabuncu M, Dalca A, Guttag J. An unsupervised learning model for deformable medical image registration. ArXiv. 2018. https://doi.org/10.1109/cvpr.2018.00964.
20. Bichu YM, Hansa I, Bichu AY, Premjani P, Flores-Mir C, Vaid NR. Applications of artificial intelligence and machine learning in orthodontics: a scoping review. Prog Orthod. 2021;22:18. https://doi.org/10.1186/s40510-021-00361-9.
21. Kim J, Kim I, Kim YJ, Kim M, Cho JH, Hong M, Kang KH, Lim SH, Kim SJ, Kim YH, Kim N, Sung SJ, Baek SH. Accuracy of automated identification of lateral cephalometric landmarks using cascade convolutional neural networks on lateral cephalograms from nationwide multi-centres. Orthod Craniofac Res. 2021;24(Suppl 2):59–67. https://doi.org/10.1111/ocr.12493.
22. Zhang H, Wang C-W, Muzakky H, Dai J, Li X, Ma C, et al. Deep learning techniques for automatic lateral X-ray cephalometric landmark detection: is the problem solved? ArXiv. 2024:2409.15834. https://doi.org/10.48550/arXiv.2409.15834.
23. Jeong SH, Yun JP, Yeom HG, Kim HK, Kim BC. Deep-learning-based detection of cranio-spinal differences between skeletal classification using cephalometric radiography. Diagnostics (Basel). 2021. https://doi.org/10.3390/diagnostics11040591.
24. Li H, Xu Y, Lei Y, Wang Q, Gao X. Automatic classification for sagittal craniofacial patterns based on different convolutional neural networks. Diagnostics. 2022. https://doi.org/10.3390/diagnostics12061359.
25. Yu HJ, Cho SR, Kim MJ, Kim WH, Kim JW, Choi J. Automated skeletal classification with lateral cephalometry based on artificial intelligence. J Dent Res. 2020;99:249–56. https://doi.org/10.1177/0022034520901715.
26. Chang Q, Bai Y, Wang S, Wang F, Liang S, Xie X. Automated orthodontic diagnosis via self-supervised learning and multi-attribute classification using lateral cephalograms. Biomed Eng Online. 2025;24:9. https://doi.org/10.1186/s12938-025-01345-0.
27. Wang CW, Huang CT, Lee JH, Li CH, Chang SW, Siao MJ, et al. A benchmark for comparison of dental radiography analysis algorithms. Med Image Anal. 2016. https://doi.org/10.1016/j.media.2016.02.004.
28. Wang CW, Huang CT, Hsieh MC, Li CH, Chang SW, Li WC, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric X-Ray images: a grand challenge. IEEE Trans Med Imaging. 2015;34:1890–900. https://doi.org/10.1109/TMI.2015.2412951.
29. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. ArXiv. 2016; abs/1611.03530. https://doi.org/10.48550/arXiv.1611.03530.
30. Gupta J, Pathak S, Kumar G. Deep learning (CNN) and transfer learning: a review. J Phys Conf Ser. 2022;2273:012029. https://doi.org/10.1088/1742-6596/2273/1/012029.
31. Wang S. Structure-aware and Texture-aware End-to-end Network. 2023. https://github
32. Lindner C, Bromiley PA, Ionita MC, Cootes TF. Robust and accurate shape model matching using random forest regression-voting. IEEE Trans Pattern Anal Mach Intell. 2015;37:1862–74. https://doi.org/10.1109/tpami.2014.2382106.
33. Arık S, Ibragimov B, Xing L. Fully automated quantitative cephalometry using convolutional neural networks. J Med Imaging (Bellingham). 2017;4:014501. https://doi.org/10.1117/1.Jmi.4.1.014501.
34. Chen R, Ma Y, Chen N, Lee D, Wang W. Cephalometric landmark detection by attentive feature pyramid fusion and Regression-Voting. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap P-T, Khan A, editors. Medical image computing and computer assisted Intervention – MICCAI 2019. Cham: Springer International Publishing; 2019. p. 873–81. https://doi.org/10.1007/978-3-030-32248-9_97.
35. Zeng M, Yan Z, Liu S, Zhou Y, Qiu L. Cascaded convolutional networks for automatic cephalometric landmark detection. Med Image Anal. 2021;68:101904. https://doi.org/10.1016/j.media.2020.101904.
36. McCouat J, Voiculescu I, Contour-Hugging Heatmaps for Landmark Detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. pp. 20565-73. https://doi.org/10.1109/CVPR52688.2022.01994.
37. Lindner C, Thiagarajah S, Wilkinson JM, The arc OC, Wallis GA, Cootes TF. Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE Trans Med Imaging. 2013;32:1462–72. https://doi.org/10.1109/TMI.2013.2258030.
38. Urschler M, Ebner T, Štern D. Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization. Med Image Anal. 2018;43:23–36. https://doi.org/10.1016/j.media.2017.09.003.
39. Anuwongnukroh N, Dechkunakorn S, Damrongsri S, Nilwarat C, Pudpong N, Radomsutthisarn W, Kangern S. Accuracy of automatic cephalometric software on landmark identification. Iop Conf. 2017;265(012028). https://doi.org/10.1088/1757-899x/265/1/012028.
40. Kim HJ, Kim KD, Kim DH. Deep convolutional neural network-based skeletal classification of cephalometric image compared with automated-tracing software. Sci Rep. 2022;12:11659. https://doi.org/10.1038/s41598-022-15856-6.
41. Steiner CC. Cephalometrics for you and me. Am J Orthod. 1953;39:729–55. https://doi.org/10.1016/0002-9416(53)90082-7.
42. Jacobson A. The wits appraisal of jaw disharmony. Am J Orthod. 1975;67:125–38. https://doi.org/10.1016/0002-9416(75)90065-2.
43. Bjork A. The face in profile. Sven Tandlak Tidskr. 1947;40. https://api.semanticscholar.org/CorpusID:202265268

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.