BioMedical Engineering OnLine

RESEARCH

Open Access

Automated orthodontic diagnosis via self-supervised learning and multi-attribute classification using lateral cephalograms

Qiao Chang[1], Yuxing Bai[1,3], Shaofeng Wang[1], Fan Wang[1], Shuang Liang[2,3,4*] and Xianju Xie[1,3*]

*Correspondence:
shliang@ccmu.edu.cn;
dentistxxj@mail.ccmu.edu.cn

[1] Department of Orthodontics, Beijing Stomatological Hospital, Capital Medical University, No. 9 Fanjiacun road, 100070 Beijing, China
[2] School of Biomedical Engineering, Capital Medical University, No. 10, Xitoutiao You An Men, 100069 Beijing, China
[3] Laboratory for Clinical Medicine, Capital Medical University, No. 10, Xitoutiao You An Men, 100069 Beijing, China
[4] Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical Application, No. 10, Xitoutiao You An Men, 100069 Beijing, China

Abstract

Background: Malocclusion, characterized by dental misalignment and improper occlusal relationships, significantly impacts oral health and daily functioning, with a global prevalence of 56%. Lateral cephalogram is a crucial diagnostic tool in orthodontic treatment, providing insights into various structural characteristics.

Methods: This study introduces a pre-training approach using multi-center lateral cephalograms for self-supervised learning, aimed at improving model generalization across diverse clinical data domains. Additionally, a multi-attribute classification network is proposed, leveraging attribute correlations to optimize parameters and enhance classification performance.

Results: Comprehensive evaluation on both public and clinical datasets showcases the superiority of the proposed framework, achieving an impressive average accuracy of 90.02%. The developed Self-supervised Pre-training and Multi-Attribute (SPMA) network achieves a best match ratio (MR) score of 71.38% and a low Hamming loss (HL) of 0.0425%, demonstrating its efficacy in orthodontic diagnosis from lateral cephalograms.

Conclusions: This work contributes significantly to advancing automated diagnostic tools in orthodontics, addressing the critical need for accurate and efficient malocclusion diagnosis. The outcomes not only improve the efficiency and accuracy of diagnosis, but also have the potential to reduce healthcare costs associated with orthodontic treatments.

Keywords: Malocclusion, Self-supervised learning, Multi-attribute classification, Lateral cephalograms, Medical image analysis

Background

Malocclusion, also known as dental misalignment, refers to the improper positioning of teeth or incorrect occlusal relationship between the upper and lower dental arches [1]. As reported by the World Dental Federation, malocclusion can significantly impact patients' daily lives, increases the risk of developing dental caries and periodontal diseases. In severe cases, it can impair essential oral functions like speech, chewing, and swallowing, potentially causing psychological health issues [2]. It is the third most

common oral health issue following dental caries and periodontal disease, with a global prevalence of 56%, which underscores the critical need for prevention and treatment of malocclusion to improve quality of life and alleviate economic burdens [3–5]. Many studies demonstrate that early diagnosis and intervention can significantly reduce the severity of future malocclusions, thereby lowering the complexity of later orthodontic treatment [6, 7].

Lateral cephalograms are widely used imaging tools for diagnosing malocclusions, treatment planning, and efficacy evaluation [8]. It provides a two-dimensional view of the skull's side profile, including the teeth, jaw, soft tissues, cervical vertebrae and airway, offering detailed insights into the craniofacial structure in a single image [9, 10]. Through the analysis of lateral cephalograms, doctors can assess the degree of skeletal and dental malocclusions in patients, enabling them to formulate appropriate treatment plans [11]. The diagnosis of skeletal malocclusions determines whether orthodontic treatment, camouflage treatment, or orthognathic surgery is necessary, while dental malocclusion diagnosis is closely related to specific treatment plans [12–14]. However, the conventional analysis process of lateral cephalograms is time-consuming, labor-intensive, and can be quite inefficient, especially facing the scenery of population screening. The diagnostic reliability of lateral cephalograms depends on the experience of dentists [15]. With the growing demand for orthodontic treatment, there is a notable shortage of qualified orthodontists, and the quality of diagnosis and treatment varies significantly across different regions [16]. This disparity greatly limits the effectiveness of lateral cephalograms as a diagnostic tool [17].

With the rapid advancement of artificial intelligence (AI), there is a growing interest in automated orthodontic diagnosis compared to manual annotation by clinicians [18, 19]. Several methods based on AI have been introduced to streamline the diagnosis process and improve efficiency in orthodontic assessments using lateral cephalograms, primarily categorized into two types: the landmark-based lateral cephalograms analysis [20] and the direct classification on lateral cephalograms [21, 22]. Automated landmark-based lateral cephalogram analysis methods hold significant utility in orthodontic diagnostics, offering efficient computational measurements against standard values for diagnostic classifications. However, these methods are susceptible to various sources of error, which can propagate through a series of calculations, making the assessment more complex and less straightforward. This error propagation can be difficult to evaluate, further complicating the reliability of the diagnostic outcomes [23]. Furthermore, in clinical measurements, using different measurement criteria may lead to contradictory diagnostic results, potentially limiting the clinical applicability of landmark-based methods [24, 25].

While the direct classification method for lateral cephalograms aims to increase diagnostic reliability by minimizing intermediate steps. Kim et al found that the direct classification model based on deep convolutional neural network was superior to automatic landmark-based method in sagittal skeletal classification [26]. Yu et al. proposed a convolution neural network with transfer learning and data augmentation techniques for single skeletal classification, with an accuracy of 90.50% [21]. Nan et al. adopted the Densenet-121 [27] network to obtain the automatic classification of the sagittal skeletal pattern in children, with the sensitivity, specificity, and accuracy of

83.99, 92.44, and 90.33%, respectively [28]. Yim et al. employed a DenseNet-169 [27] network as the classifier, and adopted the gradient-weighted class activation mapping to visualize the extracted features for automated orthodontic diagnosis, with the mean accuracy of 90.34% [29]. Li et al compared the performance on the classification of sagittal skeletal patterns using four different type of convolution neural network including Visual Geometry Group (VGG) [30], GoogLeNet [31], Residual networks (ResNet) [32], and DenseNet161 [27], with a best accuracy of 89.58% [33]. The above studies only included 1–3 classifications, which is difficult to meet the clinical needs. Chang et al. extended the diagnostic classifications to eight categories by using the DenseNet-121 network [27]. The accuracy of five diagnostic classifications were 80–90%,and the accuracy of three classifications were 70–80%, which needs to be further improved [34].

Despite these advancements, existing direct classification methods often encounter performance biases due to imbalanced sample distributions among different attributes or classes in lateral cephalograms, which is a common issue in clinical settings. Moreover, most existing methods primarily concentrate on single-attribute classification, addressing specific orthodontic diagnostic requirements. However, the craniofacial structures generally exhibit a compensatory relationship, and there are potential correlations between different attributes or classes for orthodontic diagnosis. Also, compared to multi-attribute classification tasks, training multiple single-attribute models results in extended training times and slow iteration updates, which limits their suitability for comprehensive orthodontic diagnosis in clinical settings. To address these challenges, this study proposes a novel deep learning framework, named SPMA network, for automated orthodontic diagnosis via self-supervised pre-training and multi-attribute classification using lateral cephalograms. A model weight initialization method based on masked image modeling is proposed. By pre-training the model on unlabeled data from multiple centers, it captures robust feature representations with cross-domain data distributions. A multi-attribute joint optimization network is designed, incorporating clinical prior knowledge to optimize multiple attribute classification tasks simultaneously, leveraging complementary information between different attributes to enhance performance. In clinical practice, orthodontists utilize lateral cephalograms to assess both skeletal and dental characteristics of patients, aiding in diagnosis and treatment planning. The proposed method incorporates eight specific indicators, which comprehensively describe these features and provide qualitative support for clinicians. The contributions of this work are summarized as follows:

1. A pre-training method based on multi-center lateral cephalograms was proposed, employing masked image modeling for self-supervised learning from diverse image domains, aiming to enhance model generalization when facing clinical data domain shifts.
2. A multi-attribute classification network was proposed that optimizes parameters effectively by incorporating prior correlations between attributes, utilizing complementary information to improve performance in multi-attribute classification.
3. Comprehensive evaluation on public and local clinical datasets demonstrated the superiority of this study over existing state-of-the-art (SOTA) methods, achieving a

mean accuracy of 0.9002 and providing a potential tool for automated orthodontic diagnosis.

Results

Evaluation metrics

For a comprehensive evaluation of the proposed SPMA framework, we employed various evaluation metrics, including the exact match ratio (MR), accuracy (Acc), and Hamming loss (HL) for the multi-attribute classification task. These metrics can be expressed using the following formulas:

MR: This is a strict metric that considers a sample prediction correct only if all attributes are predicted correctly. Assuming we have n samples, where y_i is the true label for the ith sample and $\hat{y}_i$ is the predicted label for the ith sample, MR can be expressed as:

$$\mathrm{MR} = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i). \tag{1}$$

Acc: This is a commonly used classification metric, representing the proportion of correctly predicted samples among the total samples. Assuming we have n samples and m attributes, where y_{ij} is the true label for the jth attribute of the ith sample and $\hat{y}_{ij}$ is the predicted label for the jth attribute of the ith sample, Acc can be expressed as:

$$\mathrm{Acc} = \frac{1}{n \times m}\sum_{i=1}^{n}\sum_{j=1}^{m} I(y_{ij} = \hat{y}_{ij}). \tag{2}$$

HL: This is a metric for multi-label classification, representing the proportion of incorrectly predicted labels among all labels. Assuming we have n samples and m attributes, where y_{ij} is the true label for the jth attribute of the ith sample and $\hat{y}_{ij}$ is the predicted label for the jth attribute of the ith sample, HL can be expressed as:

$$\mathrm{HL} = \frac{1}{n \times m}\sum_{i=1}^{n}\sum_{j=1}^{m} I(y_{ij} \neq \hat{y}_{ij}), \tag{3}$$

where $I(\cdot)$ is the indicator function, which takes the value of 1 when the condition inside the parentheses is satisfied, and 0 otherwise.

Experimental results

In this study, we conducted a series of experiments to validate the effectiveness of each component of our proposed multi-attribute classification network, the SPMA network. The SPMA network consists of a Vision Transformer (ViT) based encoder and a multi-head task network, specifically designed for automated orthodontic diagnosis. The training process of our model is divided into two stages to avoid redundancy. In the first stage, we use a self-supervised learning approach called masked image modeling for image reconstruction tasks. This process allows us to obtain pre-trained weights for the encoder. This stage lasts for 400 epochs, with a base learning rate set at 1.5e−4 and a warmup learning rate set at 1e−6. We employ a cosine scheduler for learning rate

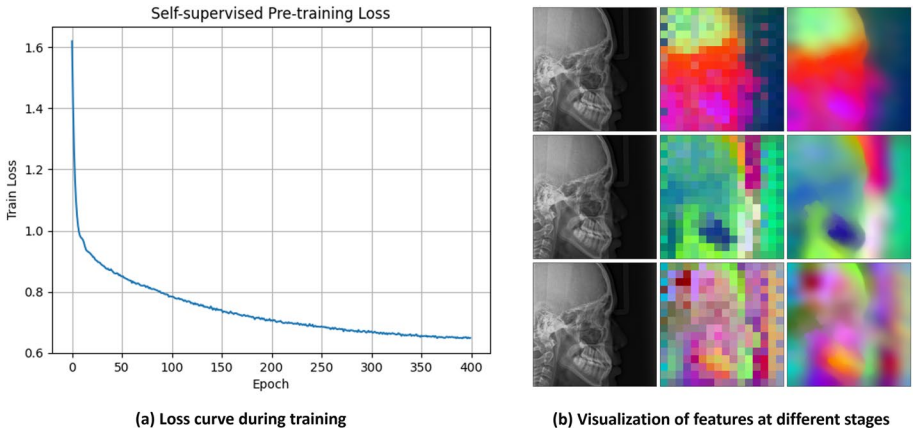

Fig. 1 The training loss (**a**) and the visualization of the extracted features (**b**)

Table 1 Multiple evaluation metrics by the SPMA network

Model	MR (%)	Acc (%)	HL (%)
SPMA	71.38	90.02	4.25

adjustment and use AdamW as the optimizer. The batch size is set to 64, and the image size is scaled to 224×224.

In the second stage, we use the encoder trained in the first stage as the feature encoder and train the entire SPMA network. This training lasts for 30 epochs, with a base learning rate of 0.001. We use a step learning rate adjustment strategy, where the learning rate is reduced by a factor of 10 every 10 epochs. The Stochastic Gradient Descent optimizer is used for training, with a batch size of 128. The images are scaled to 224×224. This two-stage training process ensures the robustness and effectiveness of our proposed SPMA network. All training process were conducted using two NVIDIA GeForce RTX 4090 GPUs((NVIDIA Corporation: Santa Clara, CA, USA).

The training loss of the self-supervised learning strategy and the visualization of the extracted features using the FeatUp module [35] are demonstrated in Fig. 1.

The results of the proposed SPMA network on multiple evaluation metrics, including MR, mean Acc, and HL, are displayed in Table 1. To the best of our knowledge, we are the first to apply multi-attribute classification in the task of automated orthodontic diagnosis using lateral cephalograms.

Ablation study

We performed ablation studies to understand the contribution of each part of our method. Specifically, we compared the performance of the encoder network obtained through self-supervised learning with that of an encoder network trained from scratch. We also conducted ablation studies on the multi-attribute classification task network and single-attribute classification task to determine the contribution of the multi-attribute joint optimization. The baseline network is consisted of a same encoder network training from scratch and single-attribute classification task network for each attribute.

Table 2 Accuracy score of each attribute obtained by the baseline and the proposed SPMA network (contributions of different part of our method were demonstrated)

Model	AP-Max (%)	AP-Mand (%)	SKFP (%)	VSFP (%)	Incl-U1 (%)	Incl-L1 (%)	AP-U1 (%)	AP-L1 (%)
Baseline	87.36	87.93	83.10	87.07	82.81	85.09	81.39	85.09
SSL	89.91	89.49	86.79	89.49	87.07	88.92	89.35	90.06
MAC	87.78	90.34	87.36	89.49	86.65	88.07	84.66	89.20
SPMA	89.35	90.62	90.62	91.34	88.35	89.91	89.06	90.91

Table 3 Accuracy score obtained by the SPMA network and the modified DenseNet, DenseNet-169 and DenseNet 121 methods

Model	AP-Max (%)	AP-Mand (%)	SKFP (%)	VSFP (%)	Incl-U1 (%)	Incl-L1 (%)	AP-U1 (%)	AP-L1 (%)
Modified DenseNet [21]	–	–	90.50	89.51	–	–	–	–
DenseNet-169 [29]	–	–	89.44	89.10	–	–	–	–
DenseNet 121 [34]	78.00	87.00	89.00	86.00	82.00	78.00	73.00	85.00
SPMA (ours)	89.35	90.62	90.62	91.34	88.35	89.91	89.06	90.91

The experimental results are presented in Table 2, where "+SSL"denotes the inclusion of the self-supervised learning strategy, and "+MAC"signifies the integration of the multi-attribute joint optimization module. And the eight attributes are maxillary anteroposterior position (AP-Max), mandibular anteroposterior position (AP-Mand), sagittal skeletal facial pattern (SKFP), vertical skeletal facial pattern (VSFP), inclination of upper incisors (Incl-U1), inclination of lower incisors (Incl-L1), anteroposterior position of upper incisors (AP-U1), and anteroposterior position of lower incisors (AP-L1).

Comparative study

Furthermore, we compared our proposed SPMA network with existing advanced automated orthodontic diagnosis methods using lateral cephalograms, including modified DenseNet [21], DenseNet-169 [29], and DenseNet 121 [34], to validate the effectiveness and advantages of our model, particularly in the context of a mixed multi-center dataset. The results of these experiments, presented in Table 3, demonstrate the superior performance of the SPMA network across various metrics. In addition, the misclassification rates were calculated and are provided in Supplementary Table 1, further supporting the evaluation of our model's performance. To visualize the activated regions during misclassifications, heatmaps are included in Supplementary Fig. 1. The attribute marked with "–" indicates that the data were not reported in the corresponding study.

For a clearer representation of the model's performance, the receiver operating characteristic (ROC) curves of the SPMA network and other SOTA methods on two metrics (SKFP and VSFP) are shown in Fig. 2. The Chi-squared test result of the ROC curve on SKFP is 129.17, with a p-value of < 0.00001, and the Chi-squared test result on VSFP is 130.71, with a p-value of < 0.00001. These results suggest a significant deviation from the null hypothesis, indicating that the SPMA model's predictions are unlikely to have

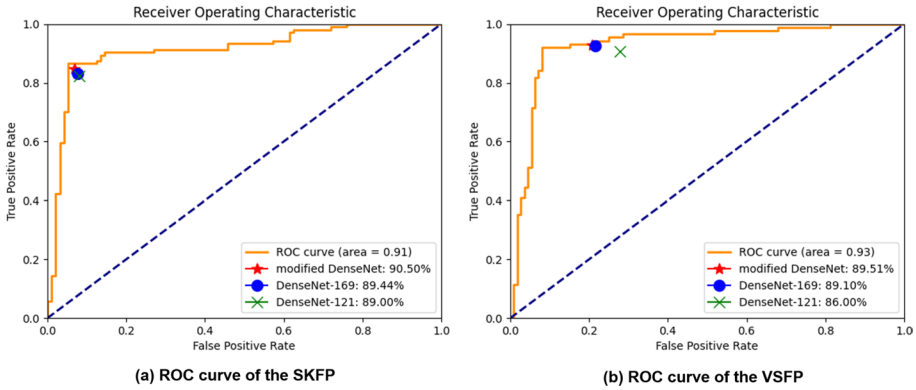

Fig. 2 ROC cures of the comparison among the proposed SPMA network and the other SOTA methods based on two selected metrics (SKFP and VSFP)

occurred by chance. This statistical significance affirms the reliability of the SPMA model's performance in classifying the data correctly. It is important to note that the Chi-squared tests were conducted specifically on the SPMA model's predictions compared to the ground truth (true labels), not directly comparing it to other models. The comparisons with other models, including the evaluation of eight parameters, are provided through quantitative performance metrics as shown in Table 3, and visual comparisons through the ROC curves are shown in Fig. 2. Based on the performance results presented above, the SPMA model demonstrates superior performance compared to the other models. However, it is important to clarify that, due to the absence of performance metrics at varying thresholds in the models from other studies, a precise statistical comparison was not available.

Discussion

The study introduces a novel deep learning framework, the SPMA network, specifically designed for automated orthodontic diagnosis using lateral cephalograms. This framework addresses several challenges in orthodontic diagnosis, such as domain shifts in clinical data and the need for effective multi-attribute classification.

One significant contribution of this work is the proposed pre-training method based on multi-center lateral cephalograms. This method leverages masked image modeling for self-supervised learning from diverse image domains. By pre-training on unlabeled data from multiple centers, the model captures robust feature representations that generalize well across different data distributions. This approach enhances the model's ability to handle domain shifts in clinical data, a common challenge in real-world orthodontic diagnosis scenarios.

Furthermore, the study introduces a multi-attribute classification network that optimizes parameters effectively by incorporating prior correlations between attributes. Clinically, while the 8 classification criteria used to describe craniofacial features are relatively independent, there are inherent relationships among them. Based on this, we introduced a multi-attribute classification network. This network architecture utilizes complementary information between different attributes, enhancing the overall performance of multi-attribute classification tasks. By jointly optimizing multiple attribute

classification tasks, the proposed network improves diagnostic accuracy and provides a more comprehensive understanding of orthodontic conditions.

The comprehensive evaluation conducted on both public and local clinical datasets demonstrates the superiority of the SPMA network over existing SOTA methods. The achieved mean accuracy of 0.9002 highlights the effectiveness of the proposed framework in automated orthodontic diagnosis. Since each classification has its own clinical significance, the aim of the study is to improve the performance of each individual classification. As shown in Table 3, compared to single-task training, the performance of each classification improved within this model. The lower performance in single-task training may be attributed to data imbalance. To achieve balanced improvement across all classifications, multi-attribute classification training for the 8 types is essential. Additionally, an error analysis was performed, revealing that most misclassifications occurred in borderline cases where the diagnostic features were less distinct. After conducting further analysis on these misclassified cases, we found that the probability values between the misclassified categories were quite close. This suggests that our model could potentially identify samples prone to confusion by incorporating a calculation of the probability difference between categories. By flagging these cases for human review, we can reduce the impact of diagnostic inaccuracies and improve overall diagnostic accuracy. These results suggest that the SPMA network has the potential to serve as a valuable tool for orthodontists, assisting them in making accurate and efficient diagnostic decisions.

Clinically, the eight indicators predicted by the proposed method comprehensively describe key craniofacial characteristics. AP-Max, AP-Mand, and SKFP reflect the sagittal development of the maxilla and mandible, as well as the relationship between them. SKFP classifications of Class II and Class III indicate the presence of skeletal deformities, necessitating more complex treatment approaches such as orthopedic correction, camouflage treatment, or orthognathic surgery compared to Class I cases. AP-Max and AP-Mand specifically illustrate the developmental status of the maxilla and mandible. The protrusion or retrusion of these structures dictates the required extraction sites and orthognathic procedures. VSFP indicates the vertical development of the jaws; severe hypodivergent or hyperdivergent cases may require orthognathic surgery. This diagnosis also influences the decision-making process for extraction plans; hyperdivergent cases generally support extraction, while hypodivergent cases require more careful consideration. Incl-U1, Incl-L1, AP-U1, and AP-L1 describe the inclination and protrusion of the upper and lower incisors, which directly affect the decision to pursue extraction-based treatment. Besides, the integration of automated orthodontic diagnosis through self-supervised pre-training and multi-attribute classification using lateral cephalograms presents substantial economic and operational benefits for public health. By reducing operation time and encapsulating the expertise of seasoned orthodontists, this approach enhances diagnostic efficiency and accuracy while minimizing errors, particularly among less experienced practitioners. Although due to its black-box nature, the underlying diagnostic logic lacks transparency, which may lead to potential misjudgments, especially in patients with ambiguous classification boundaries.The multi-attribute analysis delivers a comprehensive evaluation, swiftly processing large volumes of influential data, which is invaluable for screening, case management, and generating rich data sources for orthodontic research. These data can support epidemiological studies

and investigations into disease mechanisms, ultimately advancing the orthodontic field. Additionally, the application of multi-attribute classification provides new insights into bridging the gap between technological advancements and clinical practice. Clinically, it has been observed that there are inherent relationships between skeletal and dental characteristics. By leveraging AI, particularly the multi-attribute classification approach, we aim to incorporate these clinical experiences and patterns to enhance classification performance. The results have indeed confirmed the effectiveness of this method. Collectively, these improvements lead to more efficient and cost-effective orthodontic care, with broader implications for public health systems.

Overall, the SPMA network offers a promising approach to automated orthodontic diagnosis, combining self-supervised pre-training with multi-attribute classification to achieve superior performance. Future research directions may include further validation on larger and more diverse datasets, exploration of additional clinical attributes, and integration of real-time diagnostic support tools based on the developed framework.

Conclusion

In conclusion, this study presents a novel deep learning framework, the SPMA network, tailored for automated orthodontic diagnosis using lateral cephalograms with a best MR score of 71.38%, an accuracy score of 90.02%, and a HL loss of 0.0425%. Through innovative strategies including masked image modeling for self-supervised pre-training and multi-attribute joint optimization, the SPMA network addresses key challenges in orthodontic diagnosis, including domain shifts in clinical data and effective integration of clinical prior knowledge. Overall, the SPMA network represents a promising innovation in orthodontics, providing an automated solution for diagnosis. It has the potential to significantly benefit both orthodontic practitioners and patients.

Methods

Dataset construction

This study constructed a new dataset comprising 3310 lateral cephalograms along with their multi-attribute classification labels. The images were retrospectively selected from lateral cephalograms obtained at Beijing Stomatological Hospital between January 2015 and December 2021. The images were acquired using a Kodak 8000C dental X-ray machine (Carestream Health, Canada) with the following parameters: voltage 80 kV, current 10 mA, and X-ray exposure time of 0.5 s. Inclusion criteria for the dataset were age greater than 14 years, while exclusion criteria included motion artifacts, facial trauma, and missing incisors. The participants' ages ranged from 14 to 55 years, with a mean age of 24.5 ± 8.3 years. Notably, the presence of third molars was not documented. The lateral cephalograms were stored in Tag Image File Format with an image resolution of 1360×1840 pixels. The dataset covered features from different skeletal and dental types. Classification labels were assigned based on 8 commonly used diagnostic criteria, including AP-Max, AP-Mand, SKFP, VSFP, Incl-U1, Incl-L1, AP-U1, and AP-L1. Clinically, these 8 criteria are usually further subdivided into 3 subcategories to represent their specific subtypes. In this study, we aim to classify these subtypes across all 8 criteria simultaneously, thus improving efficiency and enhancing complementary information between the indicators to ultimately improve the model's learning performance.

These classifications were derived from a comprehensive analysis of 8 cephalometric measurement items, summarized using the Steiner analysis [36] and the Tweed analysis [37]. The specific measurements included SNA, SNB, ANB, SN-GoGn, U1-SN, IMPA, U1-NA, and L1-NB. Additionally, 324 lateral cephalograms from the publicly available 2015 Institute of Electrical and Electronics Engineers (IEEE) International Symposium on Biomedical Imaging challenge dataset were selected based on the study's inclusion criteria to construct a multi-center dataset. Two orthodontists with 8 and 5 years of experience manually measured the craniofacial features of the two datasets, and consensus labels were obtained. To ensure that the orthodontists' assessments were not biased, both were blinded to each other's measurements and to any prior patient information, allowing for independent evaluations. These two datasets together form a mixed multi-center dataset used for the performance evaluation of the methods in this study. Figure 3 displays example images from the two datasets, illustrating their distinctions. The datasets are divided into training, validation, and testing sets at a ratio of 7:2:1. The detailed information about the data distribution is presented in Table 4. This combined dataset provides a comprehensive and diverse set of data, enhancing the robustness and generalizability of the study's findings.

Data augmentation

In consideration of the significant role of geometric information in orthodontic diagnosis within lateral cephalograms, four image augmentation techniques were applied to lateral cephalograms to expand the data scale without altering the geometric information in the images. As shown in Fig. 4, these data augmentation techniques include random rotation by 10 degrees (Fig. 4a), color jittering with a brightness shift of 0.2, contrast shift of 0.2, saturation shift of 0.2, and hue shift of 0.1 (Fig. 4b), random affine transformation with a translation of 0.1 in both x and y directions (Fig. 4c), Gaussian blur with

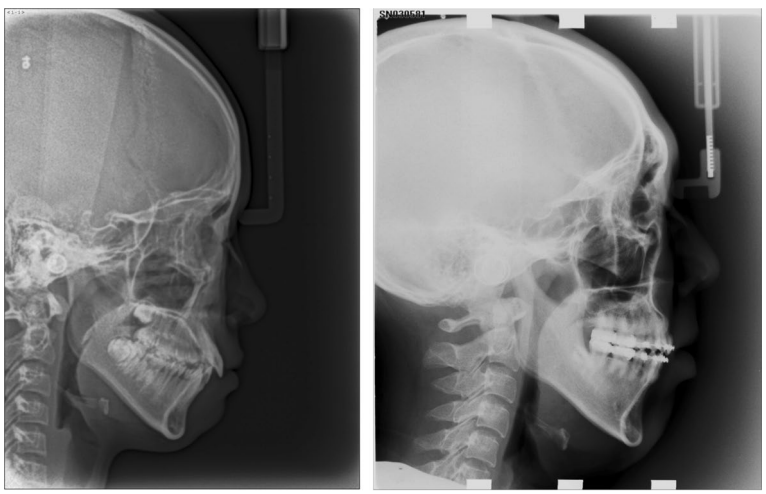

Fig. 3 Example images from the two datasets. **a** Hospital data; **b** ISBI data

Table 4 The data distribution of the hybrid multi-center lateral cephalograms dataset

Category	Type 1		Type 2		Type 3	
	Subcategory	Numbers	Subcategory	Numbers	Subcategory	Numbers
AP-Max	Retrognathic maxilla	652	Normal maxilla	2272	Prognathic maxilla	710
AP-Mand	Retrognathic mandible	788	Normal mandible	2193	Prognathic mandible	653
SKFP	Class I	1128	Class II	1404	Class III	1102
VSFP	Hypodivergent	917	Normodivergent	1811	Hyperdivergent	906
Incl-U1	Lingual inclination	628	Normal inclination	1605	Labial inclination	1401
Incl-L1	Lingual inclination	825	Normal inclination	1654	Labial inclination	1155
AP-U1	Retrusion	586	Normal	2030	Protrusion	1018
AP-L1	Retrusion	1176	Normal	1603	Protrusion	855

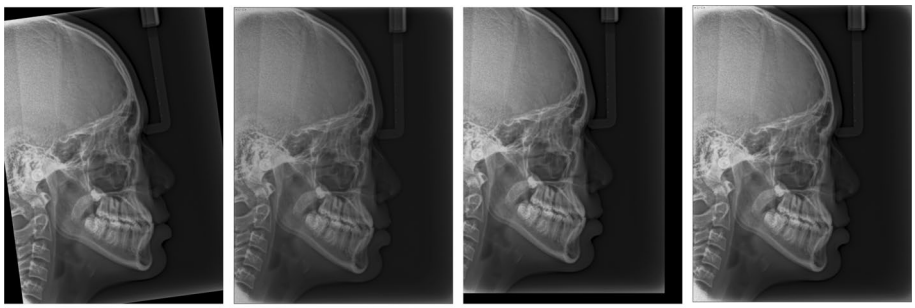

Fig. 4 Examples of data augmentation. **a** Random rotation; **b** color jittering; **c** random affine; **d** Gaussian blur

a kernel size of 3 (Fig. 4d). By employing these transformations, we aimed to reduce the potential bias and performance issues caused by imbalanced data distribution.

The pipeline of the SPMA framework

The proposed SPMA network comprises a ViT-based encoder and a multi-head task network for automated orthodontic diagnosis. The encoder initializes its weights based on a self-supervised learning task of image reconstruction. The multi-head task network achieves classification of various attributes in orthodontic diagnosis through joint optimization using multiple fully connected layers tailored for different attributes. The pipeline of the SPMA network is illustrated in Fig. 5.

Self-supervised pretraining using masked image modeling

To obtain a category-independent cross-domain feature representation, we propose a self-supervised learning-based image reconstruction method which aims to learn feature representations from unlabeled image data. The proposed self-supervised pre-training process is shown in Fig. 6.

Initially, we applied a random mask with a mask ratio of 0.75 to the input image, generating a partially masked image which is subsequently divided into several patches.

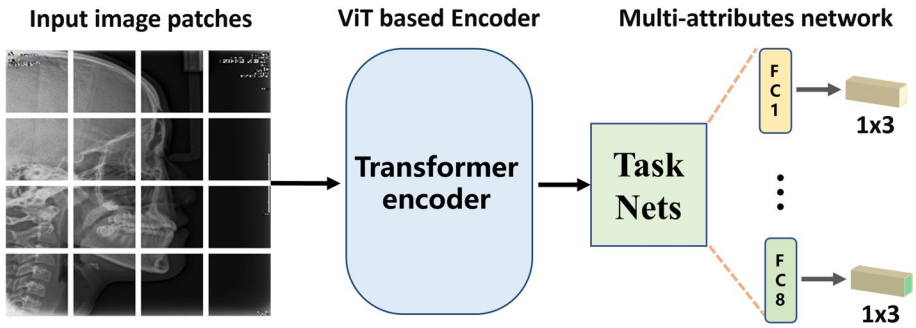

Fig. 5 The pipeline of the proposed SPMA framework. The transformer encoder took the input image patches as input and generated features for the downstream classification tasks

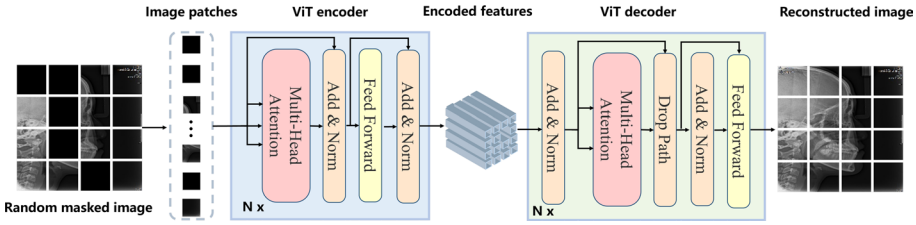

Fig. 6 The illustration of the self-supervised pre-training process. Each image was masked randomly and passed to the ViT encoder to generate the encoded features for image reconstruction

This masking and patching strategy is employed to encourage the model to focus on various parts of the image and to learn robust, spatially diverse features, which is crucial for capturing the underlying structure and relationships within the image. By dividing the image into patches, the model can analyze and reconstruct different regions independently, leading to a more comprehensive and generalized feature representation.

To be more specific, let I denote the input image. We use the PatchEmbed module to divide I into N image blocks, each of size $P \times P$. We then embed each image block into a D-dimensional vector, where D represents the embedding dimension (*embed_dim*). This can be expressed as:

$$X = \text{PatchEmbed}(I) \in \mathbb{R}^{N \times D}. \tag{4}$$

Next, we add positional embeddings (*pos_embed*) to the embedded image blocks, resulting in:

$$X = X + \text{pos\_embed} \in \mathbb{R}^{N \times D}. \tag{5}$$

Here, $\mathbb{R}$ represents the set of real numbers, and $\times$ denotes the Cartesian product. The *PatchEmbed* function maps the input image I to a matrix X with dimensions $N \times D$, where N is the number of image blocks and D is the embedding dimension. Incorporating positional embeddings enriches the embedded features with spatial information, thereby enhancing the model's representation capabilities.

These patches and their positional embeddings are then input into the vision transformer encoder, which is composed of multi-head attention and feedforward

neural networks, with add and norm operations applied sequentially. The encoded features are compactly represented and then passed through the ViT-based decoder. Specifically, we pass X through a series of Transformer blocks. Each Transformer block comprises a multi-head attention mechanism and a feedforward network, for each Block_i, the process can be represented as:

$$X = \text{Block}_i(X). \tag{6}$$

Finally, we apply a normalization layer and a linear classifier to X, resulting in the final output Y:

$$Y = \text{head}(\text{norm}(X)), \tag{7}$$

where Block_i represents the ith Transformer block, and depth signifies the total number of Transformer blocks in the series. The head function denotes the linear classifier, and norm refers to the normalization layer applied to X. This process concludes the transformation of X through the Transformer architecture, producing the output Y with enhanced features suitable for the image reconstruction task.

The output Y of the encoder was passed through the ViT decoder. The decoder, similar to the encoder, consists of multi-head attention and feedforward networks, yet includes a drop path module to enhance the stability and performance of the training process. Finally, the reconstructed image is obtained after processing through the ViT decoder. The reconstruction image exhibits enhanced details and reduced artifacts compared to the original masked image.

Multi-attribute classification network

After thorough training of the self-supervised learning model, the weights of its encoder part are saved in this study. Based on this encoder, features are extracted to construct a multi-attribute classification network. In this network, the input features Y, derived from the pre-trained encoder weights via self-supervised learning, serve as shared features. These shared features are processed by a network comprising multiple groups of fully connected layers, with each group corresponding to a specific attribute. The output for each attribute is generated as a classification output.

We denote the fully connected layer corresponding to the ith attribute as f_i. The classification output for the ith attribute, denoted as C_i, is then given by:

$$C_i = f_i(Y), \tag{8}$$

where Y represents the input features, and f_i represents the fully connected layer corresponding to the ith attribute. The classification output C_i is the output corresponding to the ith attribute.

The proposed multi-attribute classification network processes the encoded features, facilitating the simultaneous generation of classification outputs for multiple attributes. This versatility enhances the network's adaptability across various scenarios, thereby bolstering its applicability in diverse contexts.

Loss functions

In this study, considering the issue of imbalanced category distributions within various attributes, we adopted Focal Loss as the loss function for intra-attribute class classification [38]. Focal Loss is designed to address class imbalance problems, and its formal expression is as follows:

Given that there are C categories, where y represents the true category and p is the model's predicted probability distribution, the Focal Loss is defined as follows:

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{9}$$

where p_t is the predicted probability for the true category y, where $p_t = p$ when $y = 1$ and $p_t = 1 - p$ when $y = 0$. α_t is a balance factor used to adjust the weights of each category, and γ is a tuning factor used to reduce the weight of simple samples and increase the weight of difficult samples.

In this study, there are a total of 8 attribute classification tasks. The Focal Loss for each attribute i is denoted as FL_i, and each attribute has a weight w_i. Therefore, the overall loss L of the network can be represented as the weighted average of the Focal Loss for each attribute, given by:

$$L = \frac{\sum_{i=1}^{m} w_i \sum_{i=1}^{m} w_i textFL_i}{\sum_{i=1}^{m} w_i}, \tag{10}$$

where m represents the total number of attributes. The weight vector w_i is defined based on the importance of different attributes as determined by dentists. This formulation calculates the weighted average of the Focal Losses for each attribute, considering their respective weights in the overall loss of the network.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12938-025-01345-0.

Supplementary file 1.

Author contributions

Conceptualization, X.X. and S.L.; methodology, S.L. and Q.C.; investigation, Q.C. and S.L.; resources, Y.B. and S.L.; data curation, Q.C., S.W., and F.W.; visualization, Q.C., F.W., and S.L.; writing—original draft preparation, Q.C. and S.L.; writing—review and editing, Q.C., S.L., Y.B., X.X.; supervision, X.X. and Y.B.; project administration, S.L., X.X. and Y.B.; funding acquisition, Y.B., X.X., and Q.C.

Funding

This research was funded by Beijing Natural Science Foundation, Grant number: L222024, L232028.; The National Natural Science Foundation of China, Grant number: 62401378; Beijing Hospitals Authority Clinical Medicine Development of special funding support, Grant number: ZLRK202330; Innovation Research Team Project of Beijing Stomatological Hospital, Capital Medical University, grant number: CXTD20223; R&D Program of Beijing Municipal Education Commission, Grant number: KM202410025021; Capital Medical University, Grant number: PYZ23030; and Beijing Stomatological Hospital, Capital Medical University Young Scientist Program, Grant number: YSP 21-09-01.

Data availability

The data presented in this study are available on request from the corresponding author.

Code availability

The code presented in this study are available on request from the corresponding author.

Materials availability

The materials presented in this study are available on request from the corresponding author.

Declarations

Ethics approval and consent to participate
The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Beijing Stomatological Hospital (protocol code: CMUSH-IRB-KJ-PJ-2022-49 and 7 December 2022. Informed consent was obtained from all subjects involved in the study.

Consent for publication
All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare that they have no competing interest.

Received: 12 April 2024 Accepted: 24 January 2025
Published online: 04 February 2025

References

1. Scala A, Auconi P, Scazzocchio M, Caldarelli G, McNamara JA, Franchi L. Using networks to understand medical data: the case of class iii malocclusions. 2012
2. Feldens CA, Santos Dullius AI, Kramer PF, Scapini A, Busato ALS, Vargas-Ferreira F. Impact of malocclusion and dentofacial anomalies on the prevalence and severity of dental caries among adolescents. Angle Orthod. 2015;85(6):1027–34.
3. Alshammari A, Almotairy N, Kumar A, Grigoriadis A. Effect of malocclusion on jaw motor function and chewing in children: a systematic review. Clin Oral Invest. 2022;26(3):2335–51.
4. Dimberg L, Arnrup K, Bondemark L. The impact of malocclusion on the quality of life among children and adolescents: a systematic review of quantitative studies. Eur J Orthod. 2015;37(3):238–47.
5. Guo L, Feng Y, Guo H-G, Liu B-W, Zhang Y. Consequences of orthodontic treatment in malocclusion patients: clinical and microbial effects in adults and children. BMC Oral Health. 2016;16:1–7.
6. Stomatologic S. Worldwide prevalence of malocclusion in the different stages of dentition: a systematic review and meta-analysis. Eur J Paediatr Dent. 2020;21:115.
7. Wendl B, Muchitsch A, Winsauer H, Walter A, Droschl H, Jakse N, Wendl M, Wendl T. Retrospective 25-year follow-up of treatment outcomes in angle class iii patients: early versus late treatment. J Orofac Orthop. 2017;78(3):201.
8. Yang J, Ling X, Lu Y, Wei M, Ding G. Cephalometric image analysis and measurement for orthognathic surgery. Med Biol Eng Comput. 2001;39:279–84.
9. Downs WB. The role of cephalometrics in orthodontic case analysis and diagnosis. Am J Orthod. 1952;38(3):162–82.
10. Broadbent BH. A new x-ray technique and its application to orthodontia. Angle Orthod. 1931;1(2):45–66.
11. Seo H, Hwang J, Jung Y-H, Lee E, Nam OH, Shin J. Deep focus approach for accurate bone age estimation from lateral cephalogram. J Dental Sci. 2023;18(1):34–43.
12. KavithaGiri NL, Mani MS, Ahamed SY, Sivaraman G. Evaluation of central obesity, increased body mass index, and its relation to oropharyngeal airway space using lateral cephalogram in risk prediction of obstructive sleep apnea. J Pharm Bioallied Sci. 2021;13(Suppl 1):549–54.
13. Steiner CC. The use of cephalometrics as an aid to planning and assessing orthodontic treatment: report of a case. Am J Orthod. 1960;46(10):721–35.
14. Devereux L, Moles D, Cunningham SJ, McKnight M. How important are lateral cephalometric radiographs in orthodontic treatment planning? Am J Orthod Dentofac Orthop. 2011;139(2):175–81.
15. Chen S-K, Chen Y-J, Yao C-CJ, Chang H-F. Enhanced speed and precision of measurement in a computer-assisted digital cephalometric analysis system. Angle Orthod. 2004;74(4):501–7.
16. Gao Y, Sun X, Yan X, Tang Z, Lai W, Long H. Orthodontic practitioners' knowledge and education demand on clear aligner therapy. Int Dent J. 2024;74(1):81–7.
17. Gliddon MJ, Xia JJ, Gateno J, Wong HT, Lasky RE, Teichgraeber JF, Jia X, Liebschner MA, Lemoine JJ. The accuracy of cephalometric tracing superimposition. J Oral Maxillofac Surg. 2006;64(2):194–202.
18. Hwang H-W, Moon J-H, Kim M-G, Donatelli RE, Lee S-J. Evaluation of automated cephalometric analysis based on the latest deep learning method. Angle Orthod. 2021;91(3):329–35.
19. Lee J, Bae S-R, Noh H-K. Commercial artificial intelligence lateral cephalometric analysis: part 1—the possibility of replacing manual landmarking with artificial intelligence service. J Clin Pediatr Dent. 2023;47(6)
20. Lindner C, Wang C-W, Huang C-T, Li C-H, Chang S-W, Cootes TF. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci Rep. 2016;6(1):33581.
21. Yu H, Cho S, Kim M, Kim W, Kim J, Choi J. Automated skeletal classification with lateral cephalometry based on artificial intelligence. J Dent Res. 2020;99(3):249–56.
22. Chang Q, Wang Z, Wang F, Dou J, Zhang Y, Bai Y. Automatic analysis of lateral cephalograms based on high-resolution net. Am J Orthod Dentofac Orthop. 2023;163(4):501–8.
23. Hong M, Kim I, Cho J-H, Kang K-H, Kim M, Kim S-J, Kim Y-J, Sung S-J, Kim YH, Lim S-H, et al. Accuracy of artificial intelligence-assisted landmark identification in serial lateral cephalograms of class iii patients who underwent orthodontic treatment and two-jaw orthognathic surgery. Korean J Orthodont. 2022;52(4):287.
24. Freeman RS. Adjusting anb angles to reflect the effect of maxillary position. Angle Orthod. 1981;51(2):162–71.
25. Jacobson A. The "wits" appraisal of jaw disharmony. Am J Orthod Dentofac Orthop. 2003;124(5):470–9.
26. Kim H-J, Kim KD, Kim D-H. Deep convolutional neural network-based skeletal classification of cephalometric image compared with automated-tracing software. Sci Rep. 2022;12(1):11659.

27. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;4700–4708
28. Nan L, Tang M, Liang B, Mo S, Kang N, Song S, Zhang X, Zeng X. Automated sagittal skeletal classification of children based on deep learning. Diagnostics. 2023;13(10):1719.
29. Yim S, Kim S, Kim I, Park J-W, Cho J-H, Hong M, Kang K-H, Kim M, Kim S-J, Kim Y-J, et al. Accuracy of one-step automated orthodontic diagnosis model using a convolutional neural network and lateral cephalogram images with different qualities obtained from nationwide multi-hospitals. Korean J Orthodont. 2022;52(1):3.
30. Karen S. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.
31. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015;1–9.
32. 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, pp. 2016;770–778.
33. Li H, Xu Y, Lei Y, Wang Q, Gao X. Automatic classification for sagittal craniofacial patterns based on different convolutional neural networks. Diagnostics. 2022;12(6):1359.
34. Chang Q, Wang S-F, Zuo F-F, Wang F, Gong B-W, Wang Y-J, Xie X-J. Automated diagnostic classification with lateral cephalograms based on deep learning network model. Chin J Stomatol. 2023;547–553
35. Fu S, Hamilton M, Brandt LE, Feldmann A, Zhang Z, Freeman WT. Featup: a model-agnostic framework for features at any resolution. In: The Twelfth International Conference on Learning Representations 2024. https://openreview.net/forum?id=GkJiNn2QDF.
36. Steiner CC. Cephalometrics for you and me. Am J Orthod. 1953;39(10):729–55.
37. Tweed CH. The frankfort-mandibular plane angle in orthodontic diagnosis, classification, treatment planning, and prognosis. Am J Orthod Oral Surg. 1946;32(4):175–230.
38. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2017;2980–2988.

Publisher's Note