Machine Learning-Based Evaluation on Craniodentofacial Morphological Harmony of Patients After Orthodontic Treatment

Xin Wang[1†], Xiaoke Zhao[2,3,4†], Guangying Song[1,5*], Jianwei Niu[2,3,4] and Tianmin Xu[1,5*]

[1]Department of Orthodontics, Peking University School and Hospital of Stomatology, Beijing, China, [2]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China, [3]Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang University, Beijing, China, [4]Hangzhou Innovation Research Institute, Beihang University, Beijing, China, [5]NHC Research Center of Engineering and Technology for Computerized Dentistry, Beijing, China

Objectives: Machine learning is increasingly being used in the medical field. Based on machine learning models, the present study aims to improve the prediction performance of craniodentofacial morphological harmony judgment after orthodontic treatment and to determine the most significant factors.

Methods: A dataset of 180 subjects was randomly selected from a large sample of 3,706 finished orthodontic cases from six top orthodontic treatment centers around China. Thirteen algorithms were used to predict the value of the cephalometric morphological harmony score of each subject and to search for the optimal model. Based on the feature importance ranking and by removing features, the regression models of machine learning (including the Adaboost, ExtraTree, XGBoost, and linear regression models) were used to predict and compare the score of harmony for each subject from the dataset with cross validations. By analyzing the prediction values, the most optimal model and the most significant cephalometric characteristics were determined.

Results: When nine features were included, the performance of the XGBoost regression model was MAE = 0.267, RMSE = 0.341, and Pearson correlation coefficient = 0.683, which indicated that the XGBoost regression model exhibited the best fitting and predicting performance for craniodentofacial morphological harmony judgment. Nine cephalometric features including L1/NB (inclination of the lower central incisors), ANB (sagittal position between the maxilla and mandible), LL-EP (distance from the point of the prominence of the lower lip to the aesthetic plane), SN/OP (inclination of the occlusal plane), SNB (sagittal position of the mandible in relation to the cranial base), U1/SN (inclination of the upper incisors to the cranial base), L1-NB (protrusion of the lower central incisors), Ns-Prn-Pos (nasal protrusion), and U1/L1 (relationship between the protrusions of the upper and lower central incisors) were revealed to significantly influence the judgment.

Conclusion: The application of the XGBoost regression model enhanced the predictive ability regarding the craniodentofacial morphological harmony evaluation by experts after

orthodontic treatment. Teeth position, teeth alignment, jaw position, and soft tissue morphology would be the most significant factors influencing the judgment. The methodology also provided guidance for the application of machine learning models to resolve medical problems characterized by limited sample size.

1 INTRODUCTION

Malocclusion has been considered to be highly prevalent and can affect oral and facial aesthetics as well as psychosocial wellbeing in the long term (Borzabadi-Farahani, 2012; Vellappally et al., 2014). It was claimed that the facial features, especially oral aesthetics, had the potential to influence self-perceived appearance, especially during the phase of life with intense social and affective interaction (Burden, 2007). Patients seeking orthodontic treatment aim to improve their dental aesthetics and facial balance (Turley, 2015; Singh et al., 2021). As a standardized method, cephalometric analysis is routinely used to investigate the interrelationship among craniofacial bony and soft tissue landmarks and is employed as a treatment planning and evaluation tool by orthodontists, based on cephalometric radiographs before and after orthodontic treatment. It is usually based on a comparison of the values obtained from certain measurements in a group of individuals with the average values from their populations, which is set as the normal or average value. The distance and angle deviations among cephalometric landmarks for patients are compared with this value to determine whether any skeletal or dental aberration exists. However, it might be misleading to practitioners that the process of orthodontic treatment is to "correct the abnormal values for each patient." Actually, facial morphology varies in the size, shape, and position of the dentoskeletal structures for each individual, and the combinations of these morphological components are extremely diverse as well. It is important to understand that the aim of orthodontic treatment is to move teeth to a physiologically stable position and to balance the relationship within morphological components (Xu, 2017).

There has been a vast array of methods of cephalometric analysis during the past decades (Steiner, 1953; Downs, 1956; Ricketts, 1961; Tweed, 1969). Each method has some merits but may not be applicable in all cases. For a beginner in the field of orthodontics, it might be difficult to choose a certain method of cephalometric analysis that performs the best combination of the landmarks in a specific case. As practical training and clinical experience accumulate, clinicians will then be gradually familiar with each method of cephalometric analysis, which is useful in the understanding of specific morphological types and deformities. Based on accumulation by analyzing tens of thousands of patients, orthodontic experts could make validating judgments about patients' morphological harmony when reading end-of-treatment cephalometric films. A panel of orthodontic experts from similar education and practicing backgrounds could reach an agreement on the perception of the harmonious relationship between the dentition and the facial configuration (Song et al., 2013, 2014).

When experts evaluate cephalometric morphological harmony, various landmarks and components of cephalometric films are concerned. However, among them, which are the most noteworthy and important features is unclear. The aim of this study is to determine the key characteristics of greatest concern when experts comprehensively value the various landmarks and components of cephalometric films. It may give researchers further ideas about how to improve the method that represents the evaluation and delivers the thoughts of experts more effectively and precisely. Solving these problems will also help beginners to obtain a thorough understanding of balanced dental, jaw, and facial relationships after orthodontic treatment and could further improve the existing evaluation system for orthodontic treatment outcomes.

Nowadays, machine learning is increasingly being used in the medical field ranging from medical image processing and the diagnosis of specific diseases to the broader tasks of decision support and outcome prediction (Hammond et al., 2001; Rubin et al., 2017; Torlay et al., 2017; Lee et al., 2018; Livne et al., 2018; Vaquerizo-Villar et al., 2018; Wang et al., 2018; Chang et al., 2019; Dinh et al., 2019; Xu et al., 2019; Suhail et al., 2020; Verma et al., 2020; You et al., 2020). However, machine learning methods are rarely applied to evaluate craniodentofacial morphological harmony after orthodontic treatment. The present study aims to predict the evaluation of orthodontic experts and focuses on predictive modeling of applications characterized by small datasets and real-numbered continuous outputs, based on machine learning models. Such tasks, in terms of predicting the evaluation of orthodontic experts, are mostly approached by using conventional multiple linear regression models, which are based on the assumptions of statistical independence of the input variables, linearity between dependent and independent variables, normality of the residuals, and the absence of endogenous variables. However, in many applications, particularly in those involving complex physiological parameters such as values of cephalometric analysis, these assumptions are often violated (Takada et al., 2000; Lee et al., 2014, 2018). This situation would necessitate more sophisticated regression models such as machine learning, in which the system can constantly update the models through new samples to improve the efficiency and accuracy of evaluation.

In order to explore the best-fitted modeling to predict the evaluation of orthodontic experts, several such systems including linear models, SVMs, decision trees, ANNs (artificial neural networks), and ensemble models are considered in the present study. We compared the abovementioned five categories of

machine learning models involving 13 algorithms and searched for the best-fitting model for further assessing craniodentofacial morphological harmony. Based on machine learning models, our study aims to improve the prediction performance of craniodentofacial morphological harmony judgement after orthodontic treatment and to determine the most significant factors that influence the craniodentofacial morphological harmony judgement by orthodontists.

2 MATERIALS AND METHODS

2.1 Quantification of the Subjective Evaluation From Orthodontic Experts

By random stratified sampling, the dataset of 180 subjects was selected from 3,706 Chinese malocclusion patients and was analyzed with two stratified samples. One stratified sample consisted of 108 subjects from the large sample of 2,383 finished orthodontic cases in six orthodontic treatment centers around China (including the Peking University School of Stomatology, the West China College of Stomatology at Sichuan University, the School of Stomatology at the Fourth Military Medical University, the Beijing Stomatology Hospital and School of Stomatology at the Capital Medical University, the Stomatology Hospital at Nanjing Medical University, and the Hospital of Stomatology at Wuhan University). The other comprised 48 subjects from another large sample of 1, 323 finished cases in the Peking University School of Stomatology and 24 overlapping subjects randomly selected from the former samples. The posttreatment lateral cephalometric X-ray images of the former samples were evaluated by a panel of 69 judges, and the latter samples were evaluated by another panel of 36 judges. Satisfactory cases were assigned a value of "1" point, acceptable cases were given "2" points, and unacceptable cases were given "3" points. For each case, the final score was the average point of all scores by the judges.

The panel of judges was recommended by the six participating treatment centers. The inclusion criteria for judges were that each had

1) an MS or Ph.D. degree in orthodontics or experience as a research supervisor of orthodontic postgraduates
2) no less than 10 years of clinical experience in orthodontics
3) the academic rank of associate professor or above

The experts who eventually participated in this study ranged in age from 40 to 60 years. Of the 69 experts on the panel, 38 were men and 31 were women; of the 36 experts on the panel, 19 were men and 17 were women.

The overlapping 24 samples were used to verify the consistency of the judges, which had a good result. Specifically, the Pearson correlation analysis showed that the two panels of judges were significantly and positively correlated (r = 0.905, $p < 0.01$), and no significant difference was found (for the paired t test, $p > 0.05$; for the intraclass correlation coefficient, ICC = 0.902). Details of the original data are shown in **Supplementary Material S1**.
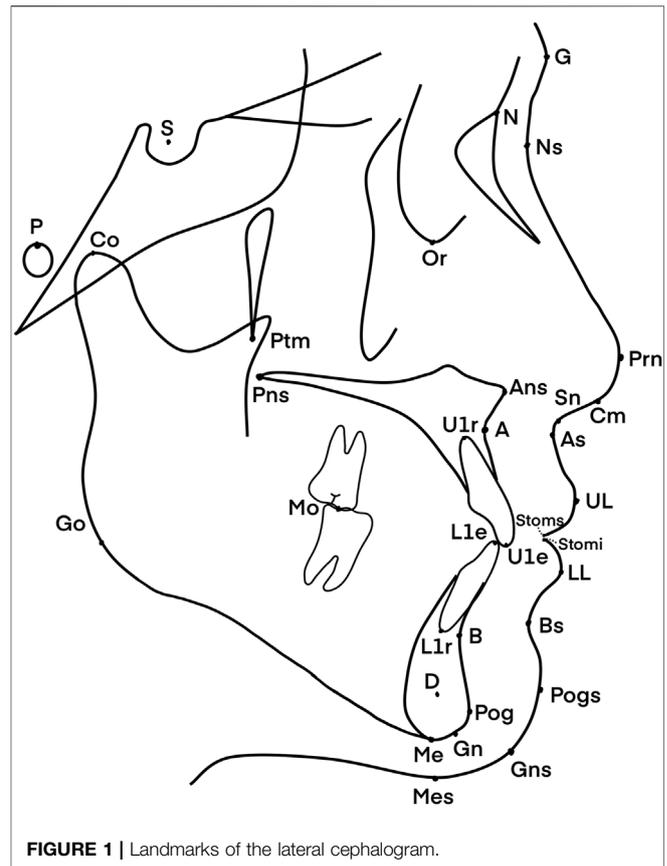

FIGURE 1 | Landmarks of the lateral cephalogram.

The whole study was performed in accordance with the Declaration of Helsinki for research involving human subjects and reviewed and approved by the Ethics and Research Committee, Peking University School and Hospital of Stomatology (PKUSSIRB-201947092).

2.2 Cephalometric Analysis

The input data under consideration were derived from the anatomy of the patients, as shown in **Figure 1**. They were based on the 42 cephalometric features, which are shown and defined in **Table 1**. The cephalometric features were measured by three practitioners who were trained at the Peking University School of Stomatology. The lateral cephalogram landmarks and cephalometric measurement items were as in **Figures 1–3** and **Table 1**.

2.3 Statistical Analysis

To evaluate the orthodontic treatment and fit it with the experts' comprehensive scoring, the following steps were taken (**Figure 4**): 1) data preprocessing; 2) feature selection and model adaption; and 3) performance evaluation.

2.3.1 Data Preprocessing

As mentioned previously, the cephalometric features and expert evaluation scores were utilized as input and output data. The first step was to make the data sets comparable. Before statistical analysis, data standardization was conducted through the z-score,

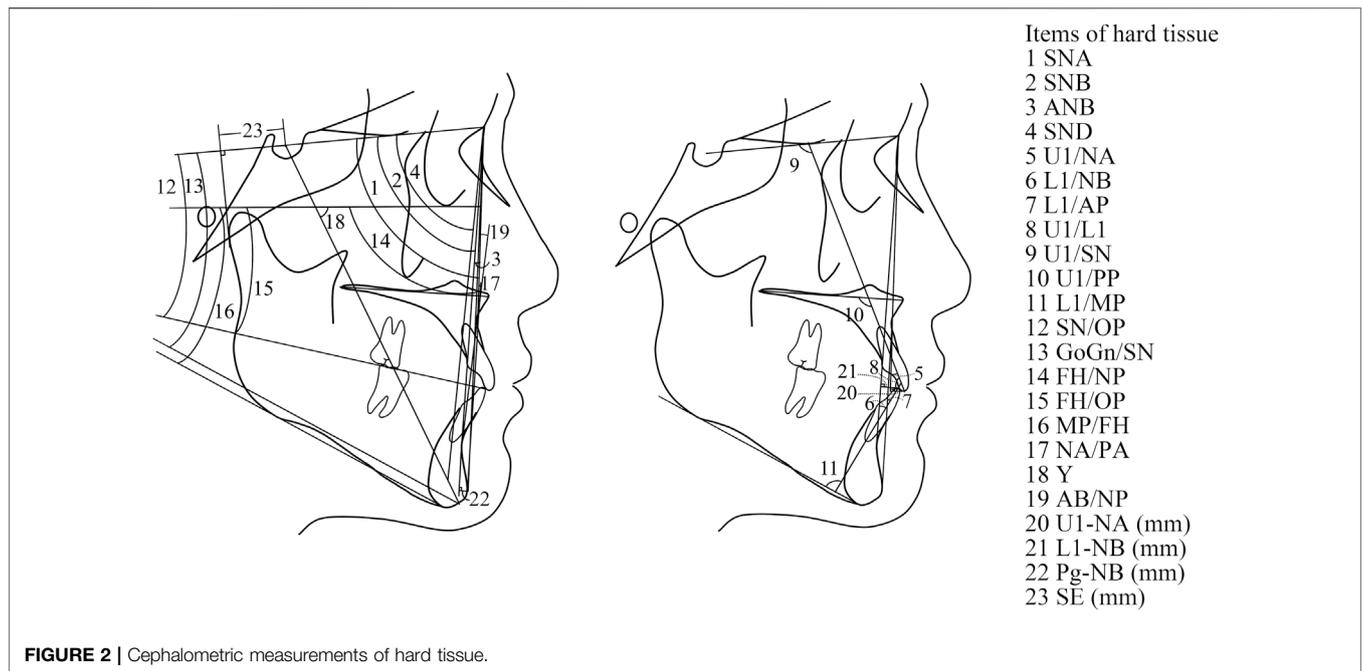**TABLE 1 |** Definitions of the 42 cephalometric features.

No.	Cephalometric variables	Definition
1	SNA	Anteroposterior position of the maxilla to the anterior cranial base (degrees)
2	SNB	Anteroposterior position of the mandible to the anterior cranial base (degrees)
3	ANB	The angle between Down's points A and B (degrees)
4	SND	The angle between the SN and ND line (degrees)
5	U1/NA	The angle between the line through the long axis of the upper central incisor and the NA line (degrees)
6	L1/NB	The angle between the line through the long axis of the lower central incisor and NB line (degrees)
7	L1/AP	The angle between the line through the long axis of the lower central incisor and the AP line (degrees)
8	U1/L1	The angle between the line through the long axis of the upper and lower central incisors (degrees)
9	U1/SN	The angle between the ling through the long axis of the upper central incisor and SN line (degrees)
10	U1/PP	The angle between the line through the long axis of the upper central incisor and palatal plane (degrees)
11	L1/MP	The angle between the line through the long axis of the lower central incisor and mandibular plane (degrees)
12	SN/OP	The angle between the SN line and occlusal plane (degrees)
13	GoGn/SN	The angle between the SN and GoGn line (degrees)
14	FH/NP	The angle between the Frankfort horizontal plane and NP line (degrees)
15	FH/OP	The angle between the Frankfort horizontal plane and occlusal Plane (degrees)
16	MP/FH	The angle between the mandibular plane and Frankfort horizontal plane (degrees)
17	NA/PA	The angle between the NA and PA line (degrees)
18	Y	Sella gnathion to the Frankfort horizontal plane (degrees)
19	AB/NP	The angle between the AB and NP line (degrees)
20	U1-NA	The perpendicular distance from U1 (incision superius) to the NA line (mm)
21	L1-NB	The perpendicular distance from L1 (incision inferius) to the NB line (mm)
22	Pg-NB	The perpendicular distance from pogonion to the NB line (mm)
23	SE	Distance between Sella and the foot point from the most posterior point of the condyle to the SN line (mm)
24	S-Ns-Sn	The angle between the S-Ns and Ns-Sn line (degrees)
25	S-Ns-Bs	The angle between the S-Ns and Ns-Bs line (degrees)
26	G-Sn-Pos	The angle between the G′-Sn and Sn-Pos line (degrees)
27	Ns-Prn-Pos	The angle between the Ns-Prn and Prn-Pos line (degrees)
28	NLA(Cm-Sn-UL)	The angle between the Cm-Sn and Sn-UL line (degrees)
29	AsUL-FH	The angle between the As-UL line and Frankfort horizontal plane (degrees)
30	BsLL-FH	The angle between the Bs-LL line and Frankfort horizontal plane (degrees)
31	AsUL-BsLL	The angle between the As-UL and Bs-LL line (degrees)
32	LL-Bs-Pos	The angle between the LL-Bs and Bs-Pos line (degrees)
33	Sn-Stoms	Distance between the subnasale and stomion superius (mm)
34	Stomi-Mes	Distance between the stomion inferius and soft tissue menton (mm)
35	Sn-Prn(FH)	The perpendicular distance from the pronasale to the line perpendicular to Frankfort horizontal plane through the subnasale (mm)
36	Ns-N(FH)	The perpendicular distance from nasion to the line perpendicular to the Frankfort horizontal plane through the soft tissue nasion (mm)
37	Sn-A (FH)	The perpendicular distance from subspinale to the line perpendicular to Frankfort horizontal plane through the subnasale (mm)
38	Bs-B(FH)	The perpendicular distance from the supramental to the line perpendicular to Frankfort horizontal plane through the most posterior point of mentolabial sulcus (mm)
39	ChinThickness	Distance between gnathion and the gnathion of soft tissue (mm)
40	UL-EP	The perpendicular distance from the upper labral to the E-line (pronaslae to pogonion of soft tissue) (mm)
41	LL-H	The perpendicular distance from the lower labral to the H-line (upper labral to pogonion of soft tissue) (mm)
42	LL-EP	The perpendicular distance from the lower labral to the E-line (pronasale to pogonion of soft tissue) (mm)

as shown in **Eq. 1**. Here, $X = \{x_i\}$ was the cephalometric feature set for all subjects, and μ and σ represented the average value and standard deviation of the normal population in China (M and X, 1965; X et al., 1986).
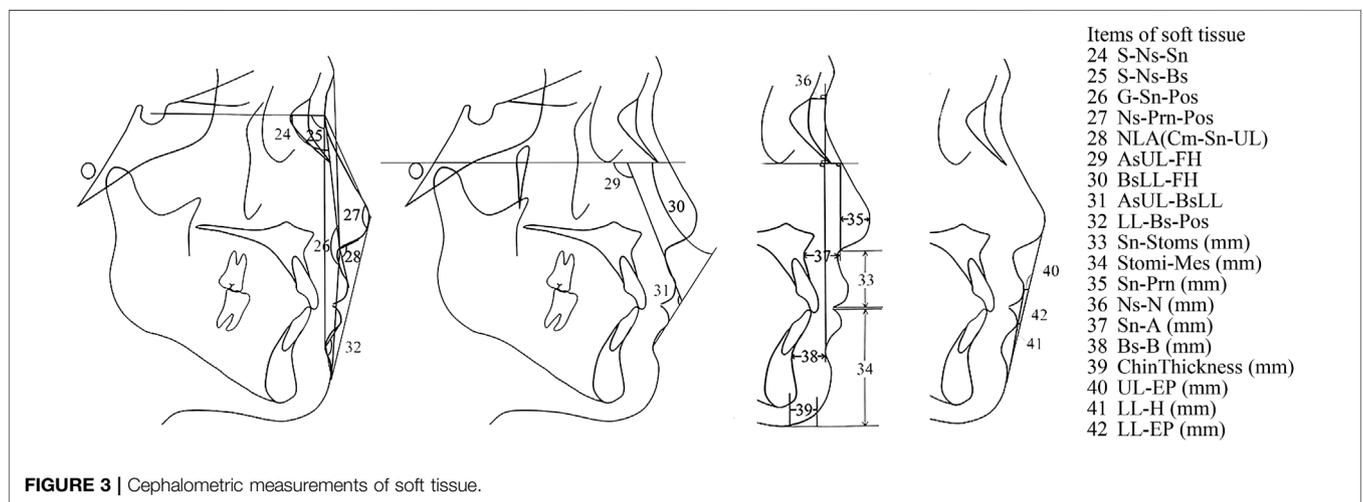
$$Z(X) = \left| (x_i - \mu)/\sigma \right|, \tag{1}$$

The data could be split into two subsets as follows: 1) a training set and 2) a testing set. When there are different settings ("hyperparameters") in an estimator, the validation set is introduced to solve the "leaking"-overfitting issues on the testing set. However, it is not suitable for our scenarios of the small sample dataset, only containing 180 samples. Splitting into
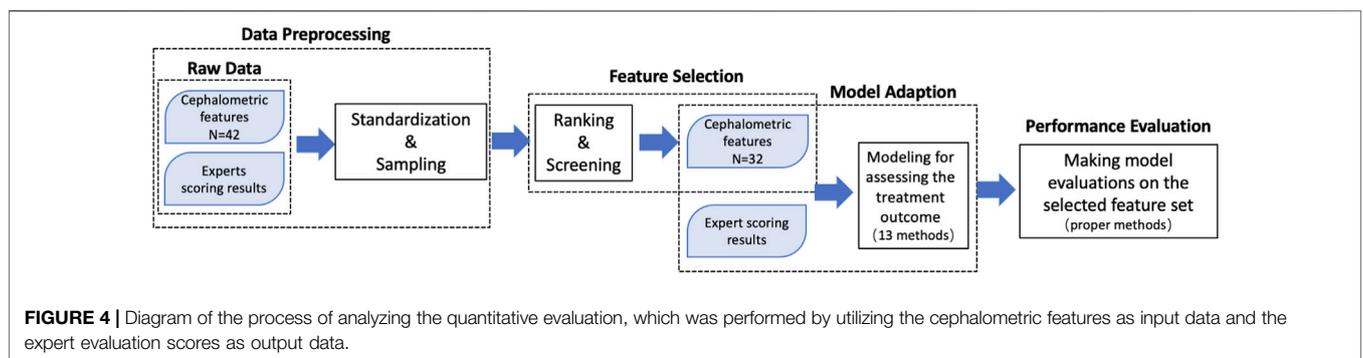
three subsets, the available training data will be further reduced for learning the model. To use the data efficiently, the procedure named cross validation (CV) is applied in our solution and the validation set is no longer needed. There are a lot of different ways to perform a CV. As shown in **Figure 5**, we have introduced two kinds of them in our solutions, which are the 10-fold procedure and the GridSearch CV with a ShuffleSplit function. We take the 10-fold procedure for evaluating per step in addition to each typical training–testing process because one evaluation on a small testing set of only 18 samples may not accurately reflect the performance of the entire model. The 10-fold procedure could allow for a fairer test as every sample has the same chance to be and to have been divided into the training set or the testing set,
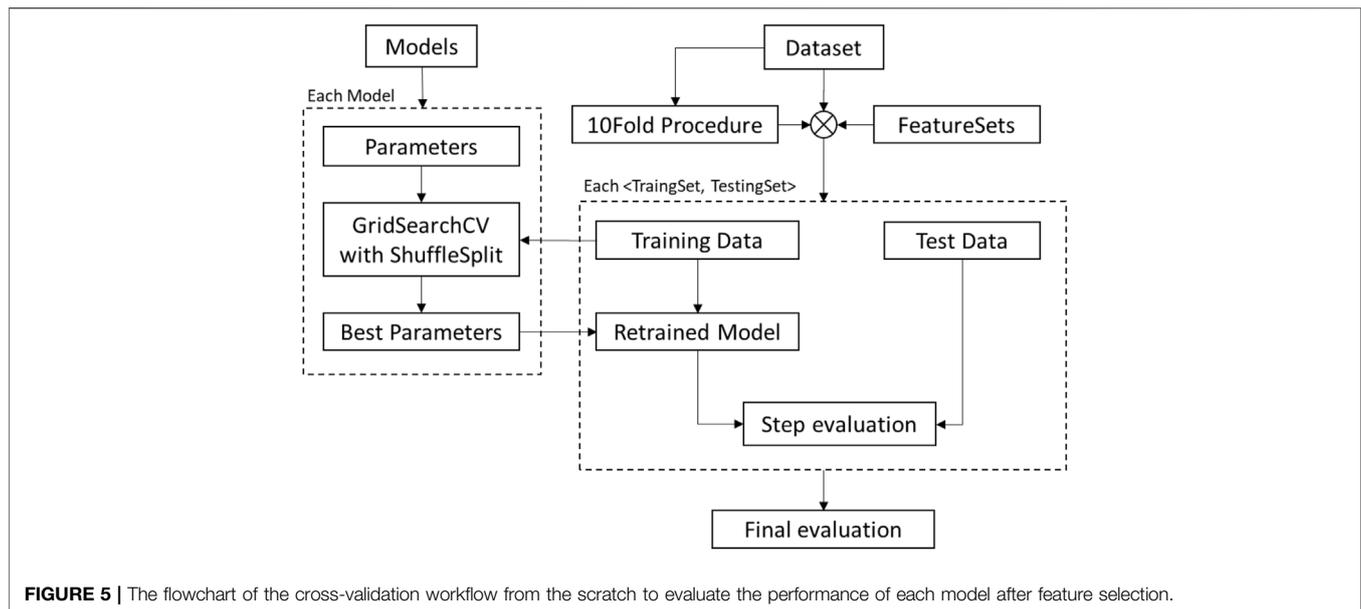
Items of hard tissue
1 SNA
2 SNB
3 ANB
4 SND
5 U1/NA
6 L1/NB
7 L1/AP
8 U1/L1
9 U1/SN
10 U1/PP
11 L1/MP
12 SN/OP
13 GoGn/SN
14 FH/NP
15 FH/OP
16 MP/FH
17 NA/PA
18 Y
19 AB/NP
20 U1-NA (mm)
21 L1-NB (mm)
22 Pg-NB (mm)
23 SE (mm)

FIGURE 2 | Cephalometric measurements of hard tissue.

Items of soft tissue
24 S-Ns-Sn
25 S-Ns-Bs
26 G-Sn-Pos
27 Ns-Prn-Pos
28 NLA(Cm-Sn-UL)
29 AsUL-FH
30 BsLL-FH
31 AsUL-BsLL
32 LL-Bs-Pos
33 Sn-Stoms (mm)
34 Stomi-Mes (mm)
35 Sn-Prn (mm)
36 Ns-N (mm)
37 Sn-A (mm)
38 Bs-B (mm)
39 ChinThickness (mm)
40 UL-EP (mm)
41 LL-H (mm)
42 LL-EP (mm)

FIGURE 3 | Cephalometric measurements of soft tissue.

FIGURE 4 | Diagram of the process of analyzing the quantitative evaluation, which was performed by utilizing the cephalometric features as input data and the expert evaluation scores as output data.

FIGURE 5 | The flowchart of the cross-validation workflow from the scratch to evaluate the performance of each model after feature selection.

and the final evaluation takes the statistical indicators on all samples at the testing set. During the GridSearch CV process, the training samples from the 10-Fold procedure are first randomly shuffled, then split into a pair of training and testing sets, and lastly, sent to select hyperparameters, train the model, and evaluate the performance of the trained model. In the ShuffleSplit function, we set it five times for the shuffle division process and take 20% of the data as the testing set.

2.3.2 Feature Selection and Model Adaption
Feature selection usually has two purposes and utilizes feature selection techniques. One technique is to reduce the clutter of original features, which includes highly correlated elements or irrelevant features. The other technique is to reduce the difficulty of analysis and increase prediction accuracy. This part of the work relies on feature engineering, which is based on the ranking of variable importance. The linear regression models often have the disadvantage of collinearity, which greatly affects error levels. In our solution, the initial feature screening was conducted to eliminate collinearity. The variable selection method was based on the correlation analysis (Ruz and Araya-Díaz, 2018) and was performed as follows:

The correlation analysis was performed on the 42 factors as well as the factors and the subjective outcomes (experts scoring results). One factor was retained out of two or more factors with Pearson correlation coefficients of 0.7 and above, with which the highest correlation with the subjective outcome was selected as the retention factor from these factors with collinearity.

Then, ten factors (SND, U1/NA, NA/PA, MP/FH, U1/PP, L1/MP, AB/NP, FH/OP, S-Ns-Sn, and LL-H) were removed, leaving 32 factors for the subsequent analysis (**Figure 6**).

For the model adaption procedure, the initial model screening was conducted to find relatively suitable algorithms from the common and widely used machine learning methods, which are listed in **Table 2**. The 13 algorithms in **Table 2** comprise five

categories of machine learning models including linear models, SVMs, decision trees, ANNs (artificial neural networks), and ensemble models, which are introduced in detail in **Supplementary Material S2**.

The mean absolute error (MAE) and root mean square error (RMSE) could be used to assess the fitting performance of the models, as shown in **Eqs 2**, **3**. Here, $X = \{x_i\}$ is the cephalometric feature set of all subjects. n stands for the total number of X, which is 180 in our case. $\{y_i\}$ is the set of ground-truth scores corresponding to each subject. $f(x_i)$ is the analytical approach that takes the cephalometric features as inputs and the prediction values as outputs.

$$MAE(X, f) = \frac{1}{n}\sum_{i=1}^{n}\left|f(x_i) - y_i\right|, \tag{2}$$

$$RMSE(X, f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - y_i\right]^2}, \tag{3}$$

2.3.3 Performance Evaluation
Based on the performance in the data processing (seen in **Figure 7**), there are four subsequent approaches: AdaBoost, ExtraTrees, XGBoost, and LSR were selected to compare subsets of cephalometric features. Here, the reason for including LSR with a mediocre performance is that LSR is utilized as a conventional approach in most medical studies. Therefore, we utilized LSR as a baseline method. We choose the other three approaches for further comparisons because they have yielded stable and relatively small metric values.

Then, the MCCV method was applied again, splitting the samples into the training set and testing set 10 times at random. The 32 factors were sequentially incorporated into the models in terms of their ranking list according to the relevance to the experts' evaluation scores. The mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient

FIGURE 6 | Pearson correlation coefficients with 42 factors.
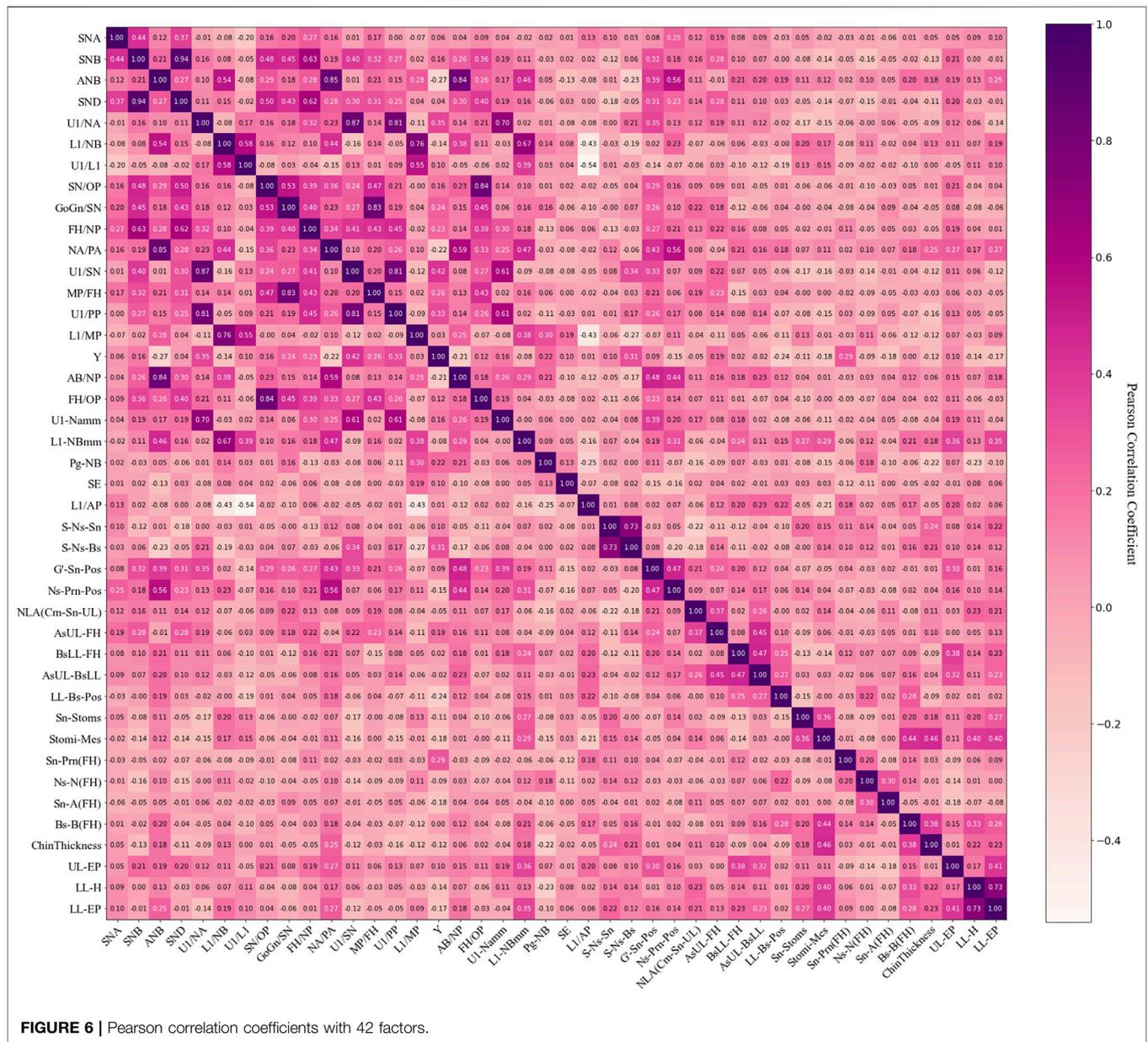
were produced to assess the final fitting performance of the models.

3 RESULTS

In this section, we evaluate the performance of four algorithms, i.e., AdaBoost, ExtraTrees, XGBoost, and LSR, by using the MCCV method. In this case, K-fold (K = 10) cross validation was applied. In the feature selection portion, we found that choosing a subset of no more than 32 features was the best option. Based on the importance rank order of these features, we gradually select feature subsets with a feature number from 1 to 32. Therefore, for each feature subset, we applied the MCCV

method and took the average values of the model computing outcomes as the final results of the metrics. We will show the results from three aspects, including the model fitting performance, predicted performance, and model interpretability.
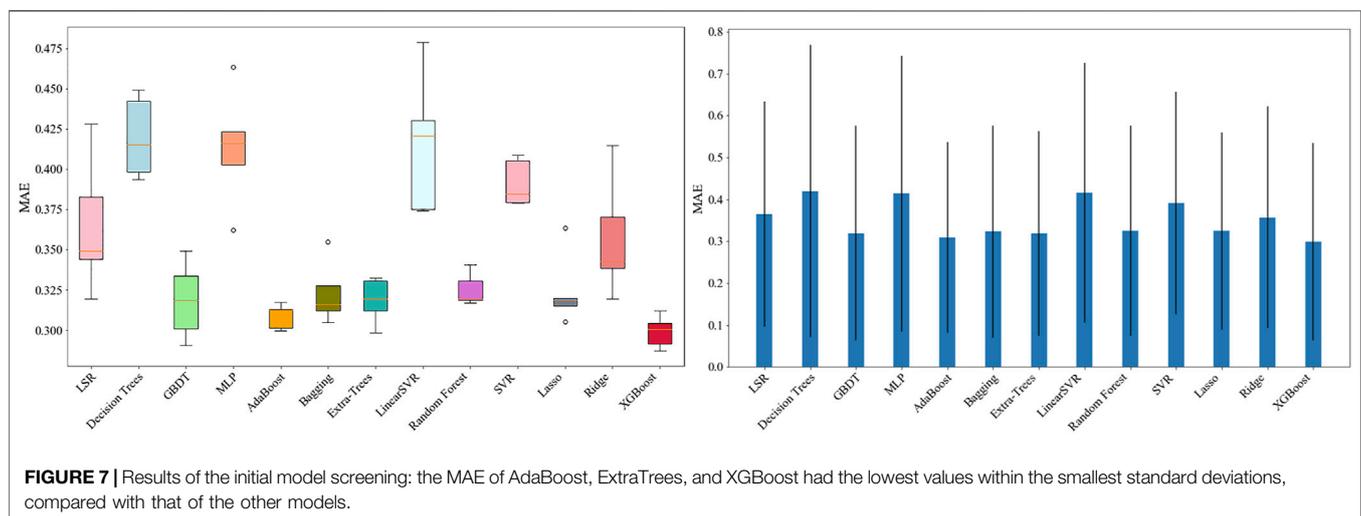
3.1 Model Fitting Performance

The line chart (**Figure 8**) could show the trend in the model fitting performance when the features (from 1 to 32) were included in turn. Such a chart could help us screen for the suitable model. In particular, when the line chart jiggles dramatically, it often means that the method is severely over-fitted. When this phenomenon occurs, the method needs to be excluded, even though it may show ideal numerical results at some nodes.

TABLE 2 | The list of 13 methods from Scikit-learn.

Name	Description
sklearn.linear_model. LinearRegression[a]	It estimates the coefficients by applying Ordinary Least Squares
sklearn.linear_model.Lasso[a]	It can estimate sparse coefficients, which addresses the issue of the least-squares penalty minimization with the ℓ_1 -norm of the coefficient vector
sklearn.linear_model.Ridge[a]	It introduces a penalty with ℓ_2-norm on the size of the coefficients to help solve the problem of collinearity
sklearn.tree. DecisionTreeRegressor[a]	It is a non-parametric supervised learning method to make predictions for a target variable by learning the decision rules inferred from the input features
sklearn. ensemble. GradientBoostingRegressor[a]	It supports a series of different loss functions. Here, we take the default loss function for regression, i.e., least squares
sklearn.ensemble. AdaBoostRegressor[a]	It assembles a sequence of weak learners with a weighted majority vote by taking the repeated boosting iteration
sklearn. ensemble. BaggingRegressor[a]	It introduces randomization into the construction procedure of an estimator and then makes an ensemble by splitting and aggregating individual predictions of this estimator on random subsets of the original training set
sklearn.ensemble. RandomForestRegressor[a]	It aims at decreasing the variance of the forest estimator by using bootstrap samples from the training set and random subsets of candidate features for node splitting
sklearn.ensemble. ExtraTreesRegressor[a]	It is similar to the random forests with node splitting. However, it randomly generates thresholds for each candidate feature and picks the best of these thresholds as the splitting rule
sklearn.neural_network. MLPRegressor[a]	It implements a multilayer perceptron (MLP) with no activation function in the output layer. Its output is a set of continuous values, and it takes the square error as the loss function
sklearn.svm.LinearSVR[a]	It is only suitable for the linear kernel when solving regression problems
sklearn.svm.SVR[a]	There are three kinds of kernels in this algorithm, i.e., linear, polynomial, and RBF kernels. Here, we take the RBF kernel
xgboost.XGBRegressor[b]	It implements the Scikit-Learn Wrapper interface for XGBoost regression

[a]*https://scikit-learn.org/stable/supervised_learning.html.*
[b]*https://xgboost.readthedocs.io/en/latest/.*

FIGURE 7 | Results of the initial model screening: the MAE of AdaBoost, ExtraTrees, and XGBoost had the lowest values within the smallest standard deviations, compared with that of the other models.

Figures 8A1–C1 show the total sample. As the number of features entering the model increased, the values of MAE and RMSE of the XGBoost regression model (the yellow line) and ExtraTrees regression model (the blue line) were less than 0.2, and the Pearson correlation coefficient was also closer to 1, which indicated a better fitting performance compared to the other two methods.
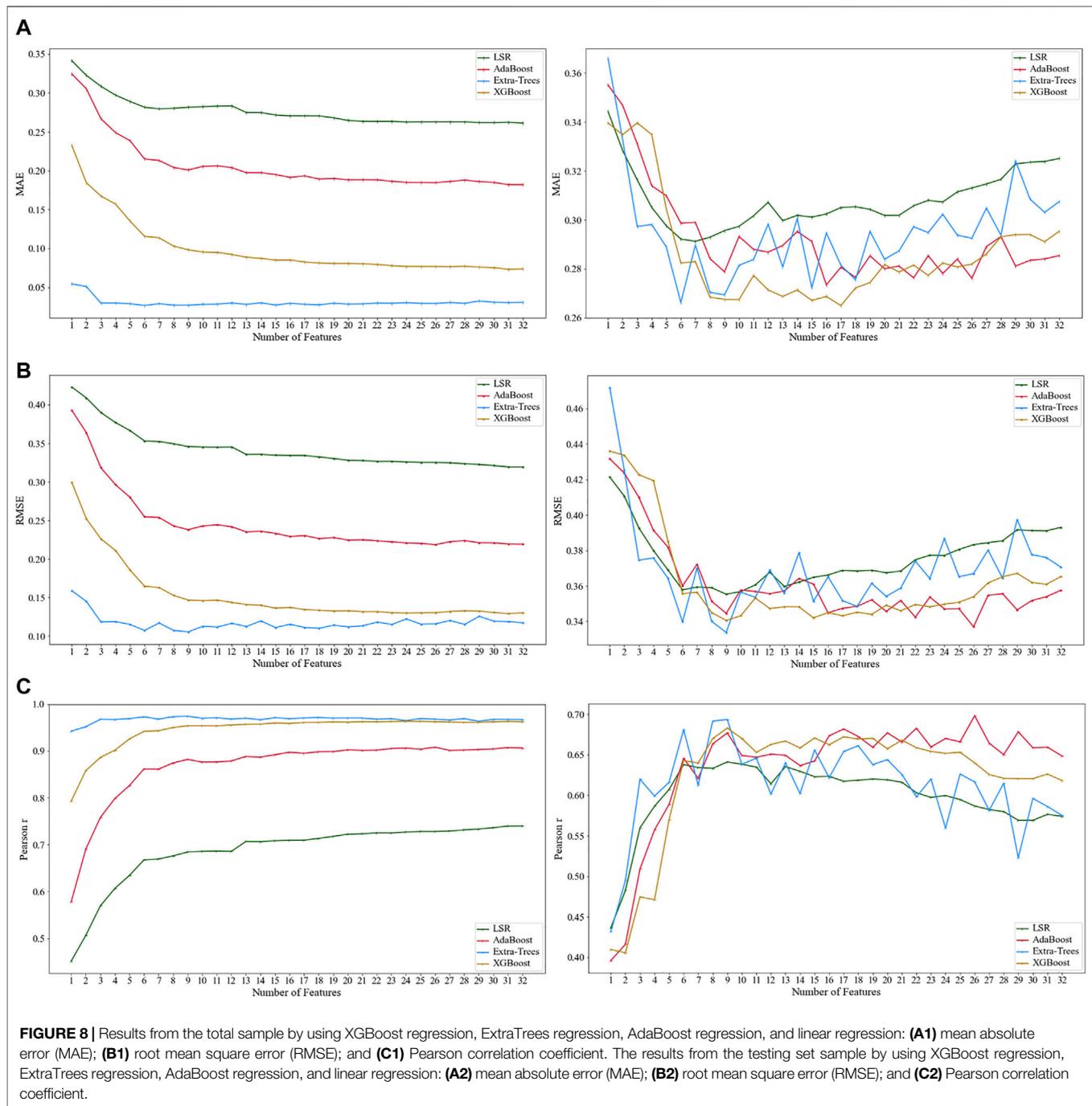
Figures 8A2–C2 show the testing set, which can more sufficiently explain the performance problem. The overall trend of the line chart showed that the MAE and RMSE values decreased at the beginning as the number of features entering the model increased and then increased as the number of factors increased after reaching the trough. In particular, the XGBoost and ExtraTrees regression models (the

yellow and blue lines) appeared more often at the troughs compared to the other two models.

However, the result of the ExtraTresss regression (the blue line) yielded a more pronounced curve jitter both in the total sample and testing set model, indicating a possible severe risk of overfitting. Therefore, the ExtraTrees regression model (the blue line) was not applicable to this study, and we ultimately concluded that the XGBoost model had better fitting performance.

3.2 Predicted Performance

To evaluate the predictive performance, the exact number of features should be clarified, when the best-predicted model is performed. Values of the MAE, RMSE, and Pearson correlation

FIGURE 8 | Results from the total sample by using XGBoost regression, ExtraTrees regression, AdaBoost regression, and linear regression: **(A1)** mean absolute error (MAE); **(B1)** root mean square error (RMSE); and **(C1)** Pearson correlation coefficient. The results from the testing set sample by using XGBoost regression, ExtraTrees regression, AdaBoost regression, and linear regression: **(A2)** mean absolute error (MAE); **(B2)** root mean square error (RMSE); and **(C2)** Pearson correlation coefficient.

coefficient from the testing set were used for comparison among XGBoost regression, ExtraTress regression, AdaBoost regression, and linear regression. In **Figure 8**, as the method of XGBoost regression was selected in Section 3.1, three nodes (when 9, 15, and 17 features were included, respectively) were then picked for further comparison, as the values at the three nodes were near the peak or trough and were relatively close to each other.

Table 3 shows the results of the values of the performance indicators from those four methods. When nine features were

included, the performance of the XGBoost regression model was MAE = 0.267, RMSE = 0.341, and Pearson correlation coefficient = 0.683; when 17 features were included, the performance of the XGBoost regression model was MAE = 0.265, RMSE = 0.343, and Pearson correlation coefficient = 0.672. Although the latter MAE value was slightly smaller than the former one, the former RMSE and Pearson correlation coefficient values were both better than those of the latter. Therefore, the XGBoost regression model exhibited the best predictive performance when nine features were included.

TABLE 3 | Machine learning model performance in the testing set.

Number of features	Performance indicator	XGBoost regression		ExtraTrees regression		AdaBoost regression		Linear regression	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
9	MAE	0.267	0.077	0.269	0.061	0.279	0.071	0.296	0.071
	RMSE	0.341	0.086	0.334	0.074	0.345	0.078	0.355	0.075
	Correlation coefficient	0.683	0.163	0.694	0.133	0.677	0.153	0.641	0.161
15	MAE	0.267	0.077	0.272	0.077	0.291	0.071	0.301	0.061
	RMSE	0.342	0.094	0.351	0.084	0.361	0.078	0.365	0.071
	Correlation coefficient	0.671	0.186	0.656	0.155	0.642	0.148	0.623	0.165
17	MAE	0.265	0.071	0.281	0.081	0.281	0.069	0.305	0.062
	RMSE	0.343	0.092	0.352	0.099	0.347	0.079	0.369	0.073
	Correlation coefficient	0.672	0.168	0.654	0.200	0.682	0.160	0.617	0.164
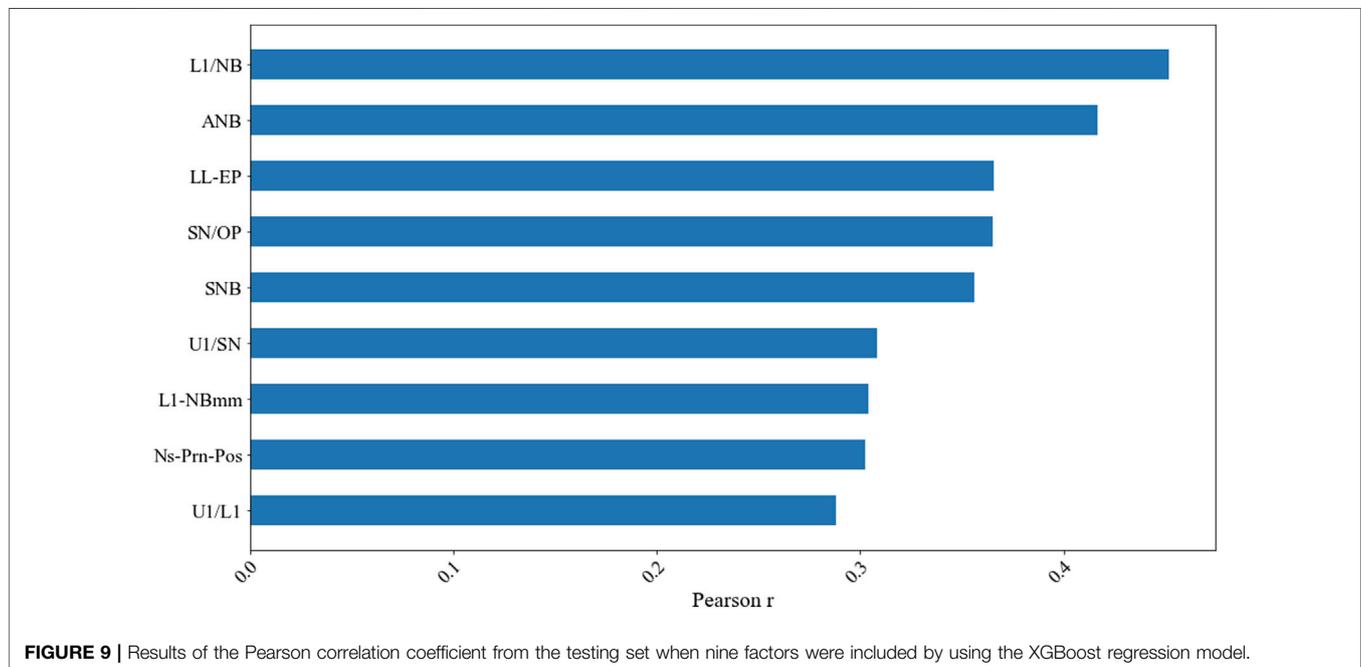
FIGURE 9 | Results of the Pearson correlation coefficient from the testing set when nine factors were included by using the XGBoost regression model.

3.3 Model Interpretability

The results of the testing set were best predicted by the XGBoost regression method when nine features were entered into the model, including L1/NB, ANB, LL-EP, SN/OP, SNB, U1/SN, L1-NBmm, Ns-Prn-Pos, and U1/L1, which reflected the lower incisors, the anterior–posterior relationship between the upper and lower jaws, the prominence of the lower lip, the steepness of the occlusal plane, the anterior–posterior position of the lower jaw relative to the cranial base, the prominence of the lower incisors, the prominence of the nose, and the relative labial inclination of the upper and lower incisors (**Figure 9**).

4 DISCUSSION

The recent advances in machine learning and a large amount of available data have laid the foundations to apply the machine learning methodology to various orthodontic aspects, including automated landmark detection on lateral cephalograms and photography images, facial attractiveness, and skeletal classification, as well as determining the degree of cervical vertebra maturation, providing orthodontic tooth extraction decisions, and predicting the need for orthodontic treatment or orthognathic surgery. Based on current studies, the most promising applications have been focused on predicting the need for treatment and decision making for tooth extractions before orthodontic treatment (Mohammad-Rahimi et al., 2021). However, the application to evaluate the craniodentofacial morphological harmony after orthodontic treatment attracts rare attention. Orthodontic treatment achieves the goal of improving the function, balance, and aesthetics of the hard and soft tissue structure by moving the teeth. Orthodontists have been working on methods to assess the results of orthodontic treatment and to be able to objectively assess the

merits of treatment results, both on a case-by-case basis and a comparison between cases. The most widely used methods of outcome evaluation include the PAR (Peer Evaluation Rating) index, the ABO-OGS (American Board of Orthodontics-Objective Grading System) evaluation system, and the ICON (Index of Complexity, Outcome and Need), each of which has its own characteristics and should be used in the evaluation of orthodontic clinical cases within a certain field. However, these methods are based on research samples and practitioners from Europe and the United States, the developed statistical methods are limited, and no orthodontic outcome evaluation system has been established for Chinese patients. The present study proposed to apply a machine learning approach to evaluate the posttreatment cephalometric diagrams of patients for Chinese orthodontic specialists, presenting a methodological innovation and an analysis of the factors incorporated into the evaluation. The model also aimed to improve the prediction performance for facial profile congruence judgment after orthodontic treatment and to find the most important characteristics when evaluating cephalometric morphological harmony by orthodontists. The model learns from past routine measurements, either including or excluding the factors concluded from the 2D images of cephalometric diagrams and/or the 3D images of plaster casts of dentitions which are used to compute the orthodontic index. The proposed XGBoost regression model was shown to be effective and precise in handling this task by performing better than the other machine learning models and traditional statistical methods that predict the scores of experts. Compared with the other approaches reported in the literature (Yu et al., 2014, 2016), the major advantages of the proposed XGBoost regression model from the present study involve the ability to deal with lower and smaller transversal data sample size.

This study identified 9 out of 42 cephalometric factors (including L1/NB, ANB, LL-EP, SN/OP, SNB, U1/SN, L1-NB, Ns-Prn-Pos, and U1/L1) that significantly influenced the orthodontist judgments when evaluating posttreatment satisfaction and facial morphological harmony. The factors could then be categorized into the following parts: the tooth position, tooth alignment, jaw position, and soft tissue morphology. These four parts cannot be separated regarding the facial morphological harmony evaluation and are structurally interlinked and influenced by each other.

For the first part, the lower incisor inclination (L1/NB), lower incisor prominence (L1-NB), upper incisor inclination (U1/SN), and relative inclination of the upper and lower incisors (U1/L1) reflect the tooth position. Some researchers (Kambara et al., 2016) investigated how the position of mandibular incisors affected facial profile aesthetics and concluded that the position of mandibular incisors for Japanese patients should be within a Holdaway ratio of 2–3 (distance from L1 to NB divided by the distance from Pog to NB) when the distance from L1 to the NB line was considered. Uesato G et al. (Uesato et al., 1978) stated that 5 mm was an ideal figure for the distance from L1 to the NB line when Steiner analysis was applied to Japanese individuals. For the second part, the occlusal plane steepness (the angle between SN and the occlusal plane, SN/OP) reflected the vertical alignment of the teeth and, to some extent, the vertical facial type. The vertical alignment of the teeth corresponded to

the occlusion of the teeth, which was determined to some extent by the vertical orientation of the facial type and the direction of the occlusal muscles. Anteroposterior and vertical facial type variations influenced the aesthetic preference of the anteroposterior lip positions and further influenced the judgement of facial harmony after orthodontic treatment. For the third part, the anteroposterior positions of the maxilla to the mandible (ANB) and the mandible to the cranial base (SNB) reflected the jaw position. For orthodontists, angle classification is the most widely used method of determining the sagittal occlusal relationship of the upper and lower teeth. Angle classification reflects to some extent skeletal malocclusion, which is the upper and lower jaw position relative to the skull. Patients with Angle Class I and Skeletal Class I (ANB = 2.7 ± 2°) usually have a normal jaw position, while patients with Angle Class II and Skeletal Class II often present with maxillary protrusion and mandibular retrusion, and patients with Angle Class III and Skeletal Class III present with maxillary retrusion and mandibular protrusion. However, with the different sagittal relationships of the upper and lower jaws, experts may develop different plans for orthodontic treatment regardless of the type of angle classification. Moreover, patients with different jaw positions may experience different difficulties and have different orthodontic treatment expectations (Turley, 2015). For the fourth part, nasal prominence/the total facial convexity angle (Ns-Prn-Pos) and lower lip prominence (lower lip to E-line, LL-EP) reflected the soft tissue morphology, which appeared to influence the aesthetics and harmony of the facial profile after orthodontic treatment. These are the features that represent the anteroposterior position of the lower lip and the amount of noise that influences the profile/facial convexity (Fortes et al., 2014). Some researchers (Perović, 2017) pointed out that the significant differences in profiles of people with class II division two compared to class I were the position of the lower and upper lip in relation to the S-line (which is another reference plane with a similar function as the E-line). Others (Joshi et al., 2015) found that the sagittal lip positions were associated with the skeletal malocclusion pattern.

The present study revealed the nine significant cephalometric features integrated into the abovementioned four parts which not only determine the most important characteristics when experts comprehensively evaluate various landmarks and components on cephalometric films but also provide evidence about the relationship among these characteristics. When assessing cases, experts may focus more on the correlation of these important factors, rather than just a standard value for a particular measurement.

Overall, the significance of our study was reflected in three main aspects: 1) first, it was the attempt to apply machine learning methods to the expert evaluation of craniodentofacial morphological harmony after orthodontic treatment, which was methodologically different from traditional statistical methods, and the results showed that the XGBoost regression model improved the fitting and predicted performance of the model over linear regression; 2) second, based on the model, we have taken the orthodontic clinical perspective to analyze the included features, which validated the content of the features of clinical interest from the machine learning perspective of our study; and 3) third, based on the first two aspects, this study provided ideas

for future exploration of similar machine learning algorithms using small samples from orthodontic clinics.

Computers and technology continue to allow us to study, predict, and produce aesthetic results that were previously thought to be unattainable. Digitalized clinical databases stored in the form of photographs, lateral cephalometric films, CBCT, 3D models, and the associated software programs have improved our ability to analyze hard and soft tissue data. In future work, further studies need to focus on exploring new solutions or enhancing the ability to utilize automation.

5 CONCLUSION

Within the limitation of the present study, the practical application of the XGBoost regression model performed a better predictive ability than that of the other models regarding the cephalometric morphological harmony evaluation by experts after orthodontic treatment. The present methodology also provided guidance for the application of machine learning models to medical problems characterized by limited datasets sizes. Moreover, the teeth position, teeth alignment, jaw position, and soft tissue morphology were demonstrated to be the most significant factors that influenced the craniodentofacial morphological harmony judgment by orthodontists.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics and Research Committee, Peking University School and Hospital of Stomatology (PKUSSIRB-201947092). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2022.862847/full#supplementary-material

REFERENCES

Borzabadi-Farahani, A. (2012). A Review of the Evidence Supporting the Aesthetic Orthodontic Treatment Need Indices. *Prog. Orthod.* 13, 304–313. doi:10.1016/j.pio.2012.03.003

Burden, D. J. (2007). Oral Health-Related Benefits of Orthodontic Treatment. *Semin. Orthod.* 13, 76–80. doi:10.1053/j.sodo.2007.03.002

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., et al. (2019). A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data. *Diagnostics* 9, 178. doi:10.3390/diagnostics9040178

Dinh, A., Miertschin, S., Young, A., and Mohanty, S. D. (2019). A Data-Driven Approach to Predicting Diabetes and Cardiovascular Disease with Machine Learning. *BMC Med. Inform. Decis. Mak* 19, 211. doi:10.1186/s12911-019-0918-5

Downs, W. B. (1956). Analysis of the Dentofacial Profile. *Angle Orthodontist* 26, 191–212. doi:10.1043/0003-3219(1956)026<0191:aotdp>2.0.co10.1111/j.1746-1561.1956.tb00805.x

Fortes, H. N. d. R., Guimarães, T. C., Belo, I. M. L., and Matta, E. N. R. d. (2014). Photometric Analysis of Esthetically Pleasant and Unpleasant Facial Profile. *Dental Press. J. Orthod.* 19, 66–75. doi:10.1590/2176-9451.19.2.066-075.oar

Fu, M., and Mao, X. (1965). X-ray Cephalometric Analysis of 144 Chinese with normal Occlusion. *J. Peking Univ. Health Sci.* 4, 251–256.

Hutton, T. J., Nelson-Moon, Z. L., Hunt, N. P., Madgwick, A. J. A., and Hammond, P. (2001). Classifying Vertical Facial Deformity Using Supervised and Unsupervised Learning. *Methods Inf. Med.* 40, 365–372. doi:10.1055/s-0038-1634194

Joshi, M., Wu, L. P., Maharjan, S., and Regmi, M. R. (2015). Sagittal Lip Positions in Different Skeletal Malocclusions: a Cephalometric Analysis. *Prog. Orthod.* 16, 8. doi:10.1186/s40510-015-0077-x

Kambara, T., Nagaya, K., Hayami, H., Kinoshita, M., Ohsaki, Y., and Kawamoto, T. (2016). Analysis of Mandibular Incisor Position in Japanese Adults : Assessment Based on the Holdaway Ratio. *J. Osaka Dent Univ.* 40, 153–158. doi:10.18905/jodu.40.2_153

Lee, H.-C., Yoon, S., Yang, S.-M., Kim, W., Ryu, H.-G., Jung, C.-W., et al. (2018). Prediction of Acute Kidney Injury after Liver Transplantation: Machine Learning Approaches vs. Logistic Regression Model. *Jcm* 7, 428. doi:10.3390/jcm7110428

Lee, H.-J., Suh, H.-Y., Lee, Y.-S., Lee, S.-J., Donatelli, R. E., Dolce, C., et al. (2014). A Better Statistical Method of Predicting Postsurgery Soft Tissue Response in Class II Patients. *Angle Orthod.* 84, 322–328. doi:10.2319/050313-338.1

Livne, M., Boldsen, J. K., Mikkelsen, I. K., Fiebach, J. B., Sobesky, J., and Mouridsen, K. (2018). Boosted Tree Model Reforms Multimodal Magnetic Resonance Imaging Infarct Prediction in Acute Stroke. *Stroke* 49, 912–918. doi:10.1161/strokeaha.117.019440

Mohammad-Rahimi, H., Nadimi, M., Rohban, M. H., Shamsoddin, E., Lee, V. Y., and Motamedian, S. R. (2021). Machine Learning and Orthodontics, Current Trends and the Future Opportunities: A Scoping Review. *Am. J. Orthod. Dentofacial Orthopedics* 160, 170–192. e4. doi:10.1016/j.ajodo.2021.02.013

Perović, T. (2017). The Influence of Class II Division 2 Malocclusions on the Harmony of the Human Face Profile. *Med. Sci. Monit.* 23, 5589–5598. doi:10.12659/msm.905453

Ricketts, R. M. (1961). Cephalometric Analysis and synthesis.Pdf. *Angle Orthodontist* 31, 141–156. doi:10.1043/0003-3219(1961)031<0141:caas>2.0.co;2

Rubin, J., Potes, C., Xu-Wilson, M., Dong, J., Rahman, A., Nguyen, H., et al. (2018). An Ensemble Boosting Model for Predicting Transfer to the Pediatric Intensive Care Unit. *Int. J. Med. Inform.* 112, 15–20. doi:10.1016/j.ijmedinf.2018.01.001

Ruz, G. A., and Araya-Díaz, P. (2018). Predicting Facial Biotypes Using Continuous Bayesian Network Classifiers. *Complexity* 2018, 1–14. doi:10.1155/2018/4075656

Singh, S., Singla, L., and Anand, T. (2021). Esthetic Considerations in Orthodontics: An Overview. *Dental J. Adv. Stud.* 9, 55–60. doi:10.1055/s-0041-1726473

Song, G.-Y., Baumrind, S., Zhao, Z.-H., Ding, Y., Bai, Y.-X., Wang, L., et al. (2013). Validation of the American Board of Orthodontics Objective Grading System for Assessing the Treatment Outcomes of Chinese Patients. *Am. J. Orthod. Dentofacial Orthopedics* 144, 391–397. doi:10.1016/j.ajodo.2013.04.018

Song, G.-Y., Zhao, Z.-H., Ding, Y., Bai, Y.-X., Wang, L., He, H., et al. (2014). Reliability Assessment and Correlation Analysis of Evaluating Orthodontic Treatment Outcome in Chinese Patients. *Int. J. Oral Sci.* 6, 50–55. doi:10.1038/ijos.2013.72

Steiner, C. C. (1953). Cephalometrics for You and Me. *Am. J. Orthod.* 39, 729–755. doi:10.1016/0002-9416(53)90082-7

Suhail, Y., Upadhyay, M., Chhibber, A., and Kshitiz, G. (2020). Machine Learning for the Diagnosis of Orthodontic Extractions: A Computational Analysis Using Ensemble Learning. *Bioengineering* 7, 55. doi:10.3390/bioengineering7020055

Takada, K., Sorihashi, Y., Stephens, C. D., and Itoh, S. (2000). An Inference Modeling of Human Visual Judgment of Sagittal Jaw-Base Relationships Based on Cephalometry: Part I. *Am. J. Orthod. Dentofacial Orthopedics* 117, 140–147. doi:10.1016/s0889-5406(00)70224-1

Torlay, L., Perrone-Bertolotti, M., Thomas, E., and Baciu, M. (2017). Machine Learning-XGBoost Analysis of Language Networks to Classify Patients with Epilepsy. *Brain Inf.* 4, 159–169. doi:10.1007/s40708-017-0065-7

Turley, P. K. (2015). Evolution of Esthetic Considerations in Orthodontics. *Am. J. Orthod. Dentofacial Orthopedics* 148, 374–379. doi:10.1016/j.ajodo.2015.06.010

Tweed, C. H. (1969). The Diagnostic Facial triangle in the Control of Treatment Objectives. *Am. J. Orthod.* 55, 651–667. doi:10.1016/0002-9416(69)90041-4

Uesato, G., Kinoshita, Z., Kawamoto, T., Koyama, I., and Nakanishi, Y. (1978). Steiner Cephalometric Norms for Japanese and Japanese-Americans. *Am. J. Orthod.* 73, 321–327. doi:10.1016/0002-9416(78)90138-0

Vaquerizo-Villar, F., Álvarez, D., Kheirandish-Gozal, L., Gutiérrez-Tobal, G. C., Barroso-García, V., Crespo, A., et al. (2018). Improving the Diagnostic Ability of Oximetry Recordings in Pediatric Sleep Apnea-Hypopnea Syndrome by Means of Multi-Class AdaBoost. *Annu. Int. Conf. Ieee Eng. Med. Biol. Soc* 2018, 167–170. doi:10.1109/embc.2018.8512264

Vellappally, S., Gardens, S. J., Al Kheraif, A.-A. A., Krishna, M., Babu, S., Hashem, M., et al. (2014). The Prevalence of Malocclusion and its Association with Dental Caries Among 12-18-Year-Old Disabled Adolescents. *Bmc Oral Health* 14, 123. doi:10.1186/1472-6831-14-123

Verma, A. K., Pal, S., and Kumar, S. (2020). Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method-A Comparative Study. *Appl. Biochem. Biotechnol.* 190, 341–359. doi:10.1007/s12010-019-03093-z

Wang, S., Li, H., Li, J., Zhang, Y., and Zou, B. (2018). Automatic Analysis of Lateral Cephalograms Based on Multiresolution Decision Tree Regression Voting. *J. Healthc. Eng.* 2018, 1–15. doi:10.1155/2018/1797502

X, Y., J, H., F, Y., and H, S. (1986). Soft-tissue Profile Analysis of 180 Chinese with normal Occlusion. *J. Clin. Stomatol* 2, 215–221.

Xu, T. (2017). *Physiologic Anchorage Control: A New Orthodontic Concept and its Clinical Application*. Springer. doi:10.1007/978-3-319-48333-7

Xu, Y., Yang, X., Huang, H., Peng, C., Ge, Y., Wu, H., et al. (2019). Extreme Gradient Boosting Model Has a Better Performance in Predicting the Risk of 90-Day Readmissions in Patients with Ischaemic Stroke. *J. Stroke Cerebrovasc. Dis.* 28, 104441. doi:10.1016/j.jstrokecerebrovasdis.2019.104441

You, W., Hao, A., Li, S., Wang, Y., and Xia, B. (2020). Deep Learning-Based Dental Plaque Detection on Primary Teeth: a Comparison with Clinical Assessments. *Bmc Oral Health* 20, 141. doi:10.1186/s12903-020-01114-6

Yu, X.-n., Bai, D., Feng, X., Liu, Y.-h., Chen, W.-j., Li, S., et al. (2016). Correlation between Cephalometric Measures and End-Of-Treatment Facial Attractiveness. *J. Craniofac. Surg.* 27, 405–409. doi:10.1097/scs.0000000000002444

Yu, X., Liu, B., Pei, Y., and Xu, T. (2014). Evaluation of Facial Attractiveness for Patients with Malocclusion: A Machine-Learning Technique Employing Procrustes. *Angle Orthod.* 84, 410–416. doi:10.2319/071513-516.1