Iran J Orthod. 2026 December; 21(2):e1446. doi:10.32592/ijorth.19.1.1446.
Published online 2026 December 1 Research Article
Original Article **Open Access**

Artificial Intelligence in Cephalometry: Machine Learning Classification of Skeletal Malocclusions in Patients Aged 6-18 years

Kazem Dalaie[1], Fatemeh Aliverdi [2], Nazila Biglar [3], Shahab Kavousinejad[2]*

[1] Associate Professor, Department of Orthodontics, School of Dentistry, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[2] Assistant Professor, Dentofacial Deformities Research Center, Research Institute of Dental Sciences, School of Dentistry, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3] Assistant Professor, Department of Orthodontics, School of Dentistry, Shahid Beheshti University of Medical Sciences, Tehran, Iran

***Corresponding author:** Shahab Kavousinejad

Address: Dentofacial Deformities Research Center, Research Institute of Dental Sciences, School of Dentistry, Shahid Beheshti University of Medical Sciences, Daneshjou Boulevard, Velenjak Street, Tehran, Iran.

E-mail: dr.shahab.k93@gmail.com

Received: 2026 June 20; **Revised:** 2026 October 27; **Accepted:** 2026 November 17

Abstract

Aim
Machine learning models offer potential for semi-automated cephalometric analysis. This study evaluated the performance of various Machine learning models in classifying sagittal (Class I, II, III) and vertical (normal, long-face, short-face) skeletal patterns in growing patients using lateral cephalometric radiographs.

Methods
A retrospective study was conducted on 1,115 lateral cephalograms (mean age 12.67 ± 3.19years; 61.1% females). Ten Machine learning models, including Artificial Neural Networks (ANN), LightGBM, and Random Forest, were trained on landmark coordinates and cephalometric indices. SMOTE was applied exclusively to training folds to address class imbalance. Model performance was evaluated through primary analysis and a "leakage-free" analysis, which excluded cephalometric indices directly used to define target labels.

Results
In the primary analysis, the ANN achieved high accuracy for vertical classification (98.96%), while the LightGBM performed best for sagittal classification. In the leakage-free analysis, where label-defining indices were excluded, performance decreased: the best-performing models were LightGBM for sagittal classification (accuracy 78.5%, macro-F1 78.5%) and ANN for vertical classification (accuracy 82.5%, macro-F1 82.0%). Key predictive features included ANB, Wits appraisal, and specific landmark coordinates (A, B, Gn, Pog).

Conclusion
Machine learning can effectively support structured cephalometric pattern recognition in growing patients. However, primary analysis performance metrics may be optimistically biased due to overlap between predictor variables and label-defining cephalometric indices. Leakage-free evaluation provides a more realistic estimate of internal predictive power. While these models show potential for clinical decision support, they currently function as consistency-support frameworks rather than independent diagnostic tools. Future validation using geographically diverse datasets and growth-aware classifiers is essential prior to clinical implementation.

Keywords: Artificial Intelligence, Cephalometry, Machine Learning, Malocclusion, Orthodontics.

Copyright © 2026, Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/) which permits copy and redistribute the material just in noncommercial usages, provided the original work is properly cited.

Background

Cephalometric analysis, introduced in the early 20th century, has evolved into an indispensable component of orthodontic diagnostics, providing quantitative insights into the spatial relationships of the maxilla, mandible, and dentition relative to the cranial base. This method involves plotting standardized landmarks on lateral radiographs to compute angular and linear measurements, such as the SNA angle for maxillary position and the ANB angle for maxillomandibular discrepancy, which are pivotal in delineating skeletal malocclusions (1,2). In pediatric orthodontics, where craniofacial growth is dynamic, cephalometrics not only helps identify deviations from normative growth trajectories but also predicts potential treatment responses, underscoring its role in preventive and interceptive care (3,4). The clinical relevance of precise cephalometric evaluation cannot be overstated, as misclassifications can result in inappropriate interventions, increased treatment duration, or even iatrogenic complications like temporomandibular disorders (5). Scientifically, it contributes to understanding genetic and environmental influences on craniofacial development and facilitates comparative studies across populations (6). However, the reliance on population-specific norms is crucial, as evidenced by variations in facial convexity and dental proclination among ethnic groups; for instance, some Asian populations have been shown to exhibit more protrusive profiles compared to Caucasian populations, necessitating customized reference values to avoid diagnostic biases (7,8).

Conventional cephalometry is affected by the subjective nature of landmark placement. Angular measurements can vary by up to 5 degrees depending on examiner experience and image quality, leading to inconsistencies in multicenter studies and routine practice (9,10). Artificial intelligence (AI)-based systems have shown promising improvements in reproducibility, with meta-analyses reporting fewer landmark-detection errors than human operators (1,11). However, gaps persist: many AI models are trained on homogeneous datasets, predominantly from Western adults, overlooking pediatric growth variations and ethnic diversity, such as in Iranian populations where normative data are sparse (6,12).

Furthermore, while sagittal classification has been extensively explored, integrated models for simultaneous sagittal and vertical assessment using diverse machine learning architectures remain underexplored, limiting their applicability in comprehensive diagnostics (13). This study was motivated by the potential of machine learning to mitigate these issues, offering a scalable solution that combines landmark coordinates with demographic and angular features for enhanced accuracy (5).

By employing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) for data balancing and hyperparameter tuning, the study sought to address class imbalance and improve model stability during training. Ultimately, this research explores the potential of machine learning to support cephalometric skeletal classification while accounting for population-specific variation (14).

So, the aim of this study was to evaluate supervised machine learning models for classifying sagittal and vertical skeletal patterns in Iranian patients aged 6–18 years using lateral cephalometric data.

By combining patient demographics, landmark coordinates, and standard measurements—while using SMOTE for data balancing and strict leakage-free testing—this research offers a clear, structured look at how well machine learning can evaluate skeletal patterns while accounting for local population differences.

Methods

Data Collection

This retrospective study evaluated sagittal and vertical skeletal relationships using lateral cephalometric radiographs of patients aged 6–18 years who attended the School of Dentistry, Shahid Beheshti University of Medical Sciences, between 2017 and 2023. The study protocol was approved by the Ethics Committee of the School of Dentistry, Shahid Beheshti University of Medical Sciences (code: IR.SBMU.DRC.REC.1403.006) and the use of anonymized archived cephalometric radiographs for secondary research and machine-learning analyses was approved by the Committee. The requirement for informed consent was waived because the study involved previously acquired, fully anonymized radiographic data.

Patient selection was guided by strict predefined eligibility criteria to ensure sample reliability and image accuracy. Individuals were included if they were between 6 and 18 years of age at the time of exposure, had archival records consisting of high-standard lateral cephalometric radiographs taken with a calibrated cephalostat to maintain consistent magnification, and

demonstrated optimal image clarity allowing reliable identification of all anatomical landmarks necessary for sagittal and vertical skeletal measurements. Conversely, patients were strictly excluded if they presented with a history of previous orthodontic treatment, orthognathic surgery, or primary dentofacial orthopedic therapy, as these interventions would alter natural craniofacial growth patterns. Additional exclusion criteria comprised any diagnosed congenital craniofacial anomalies, developmental syndromes, or pathosis(including cleft lip and/or palate, hemifacial microsomia, or craniosynostosis) to maintain sample homogeneity and eliminate confounding skeletal variations. Furthermore, radiographs were systematically excluded if they exhibited poor diagnostic quality, such as severe motion artifacts, inadequate soft or hard-tissue contrast, unreadable anatomical structures, or improper patient positioning during image acquisition including significant head rotation, tilting, or incorrect natural head posture. Technical errors such as images captured without a cephalostat or missing exposure parameters also warranted immediate disqualification. In cases where multiple longitudinal or repeated radiographs were available for the same subject, duplicate records were removed, retaining only a single representative, highest-quality radiograph per patient according to the predefined selection protocol. Following the thorough screening of initial records by the research team for image quality and compliance with all criteria, a final total of 1,115 radiographs were included in the statistical analysis.

Landmark Identification and Cephalometric Features

Following image selection, two calibrated examiners independently identified cephalometric landmarks and calculated sagittal and vertical indices based on standardized definitions. Both examiners were trained under the supervision of experienced orthodontic specialists with at least 20 years of clinical experience. In cases of uncertainty or borderline classification, final decisions were reviewed by senior orthodontists. Borderline cases were operationally defined as cases in which the available cephalometric measurements did not provide a clear classification according to the predefined normative ranges or yielded discordant indications between the primary and complementary cephalometric criteria.

The identified landmarks included: S (Sella), Se(Sella enterance), N (Nasion), Or (Orbitale), Po (Porion), A (Point A), B (Point B), Pog (Pogonion), Me (Menton), Gn (Gnathion), Go (Gonion), Co (Condylion), upper and lower incisor tips and apices, ANS (Anterior Nasal Spine), PNS (Posterior Nasal Spine), Ar (Articulare) and PTM (Pterygomaxill fissure.)

The cephalometric angular and linear measurements included: SNA, assessing the anteroposterior position of the maxilla relative to the cranial base; SNB, evaluating the anteroposterior position of the mandible relative to the cranial base; ANB, determining the sagittal relationship between the maxilla and mandible; mandibular length (Co-Pog); maxillary length (PNS-ANS); difference in jaw length, evaluating the discrepancy between maxillary and mandibular lengths; Point A to nasion perpendicular, measuring the linear skeletal position of the maxilla; Pog to nasion perpendicular, assessing chin prominence; maxillary plane to SN, evaluating the inclination of the palatal plane; Wits appraisal, assessing sagittal jaw discrepancy by projecting points A and B onto the occlusal plane; facial angle (N-Pogt to Frankfort Horizontal plan(FH)), indicating facial prognathism or retrognathism; basal angle (palatal plane to GoMe), evaluating vertical facial proportions; upper (N-ANS) and lower anterior facial heights (ANS- Me); anterior facial height ratio, Calculated as the ratio between UAFH and total anterior facial height (N-Me); saddle angle (N–S–Ar), evaluating cranial base flexure; articular angle (S–Ar–Co/Go), assessing the relationship between the cranial base and mandibular ramus; gonial angle (Ar–Go–Me), representing mandibular morphology and growth pattern; sum of posterior angles, combining cranial base and mandibular angles; SN–GoMe and SN–GoGn, assessing mandibular plane inclination; FMA and FMIA in Tweed triangle; Y-axis (SGn-FH), indicating facial growth direction; S_Ar, measuring posterior cranial base length; S-Go to N-Me (Jaraback Index), determining vertical growth patterns; ramus height (Ar-Go) , measuring vertical dimension of the mandibular ramus; U1 to NA distance and angle, evaluating maxillary incisor position and inclination; U1 to FH and U1 to palatal plan, measuring upper incisor orientation; L1 to NB distance and IMPA, evaluating lower incisor protrusion and inclination; interincisal angle, measuring the relative angle between upper and lower incisors; A_B diff, measuring horizontal skeletal discrepancy; FH to SN (SN validity); occlusal plane-SN, measuring occlusal plane inclination; SN-Pog, evaluating chin position relative to the cranial base; Se-N, measuring anterior cranial base length; and ramus length (Co-Go), evaluating ramus dimensions.

These quantitative features, along with inclination angles representing the tilt of structures relative to reference planes, were combined with the X and Y landmark coordinates to construct the comprehensive input feature space for machine learning-based classification.

Feature Selection and Reliability Assessment

During exploratory analysis, pairs of candidate predictors with strong correlations ($|r| \geq 0.80$) were identified to assess redundancy and potential multicollinearity. Correlation-based feature selection was performed using the training data only. For each highly correlated pair, one redundant predictor was excluded according to the predefined feature-selection procedure, while the remaining predictor was retained based on its clinical relevance and predefined role in the classification task. Separate predictor sets were specified for sagittal and vertical skeletal classification. The candidate predictors included age and sex, together with sagittal skeletal parameters (SNA, SNB, ANB, and Wits appraisal) and vertical skeletal parameters, including the sum of posterior angles, occlusal plane angle, maxillary plane angle, mandibular plane angle, and basal plane angle. The variables excluded during the correlation-based screening and the final predictor sets for each classification task are reported in Table 1. No correlation-based feature selection or other data-driven preprocessing was performed using the independent test set. Multiple calibration sessions were conducted before the main analysis to assess measurement reliability. Intra-examiner and inter-examiner reliability were evaluated using intraclass correlation coefficients (ICCs). The ICC model, type of agreement, and corresponding 95%confidence intervals were reported for each measurement. Most variables demonstrated good-to-excellent reliability, with ICC values exceeding 0.80.

Landmarking and Dataset Preparation

The radiographs were randomly allocated to the two examiners using a computer-generated random sequence. Cephalometric landmark identification and measurements were subsequently performed using Dolphin Imaging software (version 10.5; Dolphin Imaging & Management Solutions, Chatsworth, CA, USA). Skeletal classifications were assigned based on sagittal and vertical indices. Figure 1 depicts the Dolphin software user interface used for cephalometric analysis.

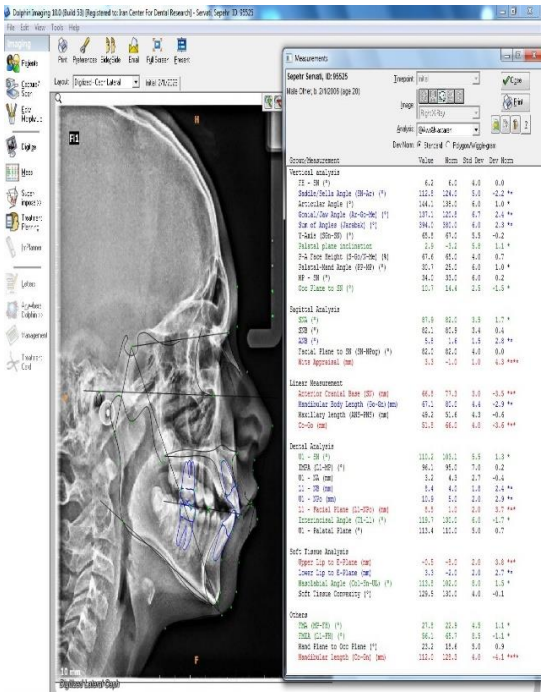

Figure 1. User interface of the Dolphin software for cephalometric analysis.

Table 1. Predictor variables retained and excluded in the primary and leakage-free machine learning configurations for sagittal and vertical skeletal classification

Task	Analysis mode	Included variables (predictors)	Excluded variables (to prevent data leakage)
Sagittal	Primary	All 49 cephalometric variables + X and Y coordinates of the 20 landmark points (age, gender, SNA, SNB, ANB, Wits, mandibular length, maxillary length, jaw-length discrepancy, Point A & Pog to nasion perp, maxillary plane to SN, facial angle, SN–Pog, PP–GoMe, UAFH, LAFH, anterior facial height ratio, SN–GoMe, FMA, Y-axis, occlusal plane–SN, SN–GoGn, saddle angle, articular angle, gonial angle, sum of posterior angles, S–Ar, Jarabak index, ramus height, U1–NA, U1–SN, U1–FH, L1–NB, interincisal angle, IMPA, FMIA, inclination angle, FH–SN, Se–N, ramus length, and specific length variants.	None
Sagittal	Leakage-free	Same as Primary, but excluding the primary diagnostic label-defining parameters.	ANB, Wits, SNA, SNB (excluded to prevent data leakage)
Vertical	Primary	All 49 cephalometric variables + X and Y coordinates of the 20 landmark points (same complete variable set as above).	None
Vertical	Leakage-free	Same as Primary, but excluding the primary vertical diagnosis-defining parameters.	SN–GoGn, SN–GoMe, Jarabak index (S–Go/N–Me), sum of posterior angles (excluded to prevent data leakage)

The feature set comprises standard cephalometric angular and linear measurements alongside the X- and Y -coordinates of the 20 anatomical landmarks (N,S,Ar, Go,Me, Gn, Pog, A, B, Or, Molar/Premolar cusp tips, U1, U1-apex, L1, L1-apex, PNS, ANS, Se, and PTM

Age- and gender-specific normative ranges were adopted according to the Iranian population standards reported by Azarbayejani et al. (15) (Table 2). Because Wits appraisal normative values were not reported in the Iranian cephalometric studies reviewed for this study, including the study by Azarbayejani et al., the classic sex-specific normative values established by Jacobson (16) were used for Wits-based classification. As these values were not age-stratified, the same sex-specific thresholds were applied across all predefined age groups.

Accordingly, samples were classified into three sagittal skeletal groups (Class I, Class II, and Class III) and Vertical skeletal patterns were categorized into three groups (normal, long-face, and short-face) using four cephalometric indicators: the SN–GoGn angle, SN–GoMe angle, Jarabak index (S–Go/N–Me ratio), and sum of posterior angles. In the anteroposterior dimension, the Wits appraisal was used as a complementary measure to ANB for resolving borderline cases but was not applied as a standalone criterion. All borderline or ambiguous cases were independently reviewed and confirmed by orthodontists. Because these adjudicated cases were determined through expert review rather than a fully independent reference standard, the adjudication process may have introduced a degree of labeling subjectivity. The final dataset included the X and Y coordinates of cephalometric landmarks, along with angular and linear cephalometric variables. Skeletal classification labels were appended as outcome variables: (a) sagittal classification (Class I, II, and III), and (b) vertical growth pattern (normal = 1, long face = 2, and short face = 3). Data were exported in Comma-separated values (CSV) file format and used as input for machine learning models. All subsequent data processing, modeling, and analyses were performed using Python Software Foundation, Wilmington, USA (Figure 2).

Table 2. Normative ranges of sagittal and vertical cephalometric indices in male and female patients used for labeling. ANB, SNA, SNB, and SN–GoGn are expressed in degrees (°), whereas Wits appraisal is expressed in millimeters (mm).

Index	Sex	Age 6–7	Age 8–9	Age 10–11	Age 12–13	Age 14–18
ANB°	Male	2.73–6.19	0.96–5.88	1.28–4.76	0.73–5.35	0.97–4.31
	Female	3.77–6.95	2.00–4.94	1.23–5.15	1.42–5.26	0.46–4.84
SNA°	Male	75.92–81.86	75.80–83.94	74.95–83.55	74.88–85.56	76.68–83.24
	Female	77.20–83.56	76.34–82.08	76.43–83.27	76.73–83.33	78.25–86.07
SNB°	Male	70.73–78.11	74.08–78.82	72.32–80.12	73.21–81.15	73.90–80.74
	Female	71.81–78.23	73.15–78.31	73.30–80.02	73.49–79.89	75.70–83.32
Wits mm	Male	-0.9–2.9	-0.9–2.9	-0.9–2.9	-0.9–2.9	-0.9–2.9
	Female	-1.77–1.77	-1.77–1.77	-1.77–1.77	-1.77–1.77	-1.77–1.77
GoGn-SN°	Male	30.21–42.67	28.05–41.91	25.08–39.18	29.05–43.61	28.28–37.42
	Female	29.32–43.88	29.32–42.88	28.80–43.34	26.40–39.90	25.89–40.17

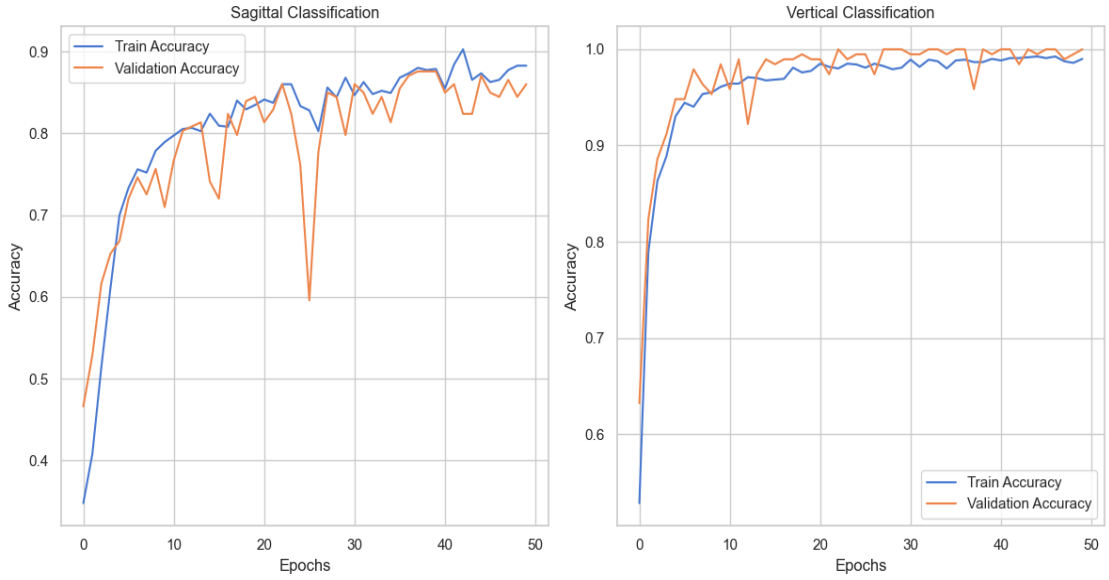

Figure 2. Training and validation accuracy and loss curves of the artificial neural network during model development for sagittal and vertical skeletal pattern classification.

Data Preprocessing

Exploratory Data Analysis

Data preprocessing included data cleaning, missing-value handling, duplicate removal, and correction of inconsistencies. The dataset was first divided into a training set (80%) and an independent test set (20%). The test set was set aside and was not used in any preprocessing, resampling, feature-selection, hyperparameter-tuning, or model-selection step.

All data-dependent preprocessing steps, namely missing-value imputation, Min–Max scaling, landmark-coordinate normalization, correlation-based feature selection, and resampling, were performed within the training set only. Within the training set, k-fold cross-validation was used for hyperparameter tuning and model selection. In each fold, every preprocessing step, including correlation-based feature selection, was fitted only on the training portion of that fold. It was then applied unchanged to the corresponding validation portion. Landmark coordinates were normalized to account for differences in image scale.

To address class imbalance, SMOTE was applied only to the training portion of each cross-validation fold. No synthetic samples were generated from validation-fold or test-set observations.

After model selection, the final model was retrained on the complete training set (80%). All

preprocessing parameters and the feature-selection rules were re-estimated from this training set only. The resulting pipeline was then applied unchanged to the independent test set, which was used once for final performance evaluation. No external validation dataset was available, so the generalizability of the findings should be interpreted with caution.

Machine Learning Models

A range of supervised machine-learning algorithms was implemented to classify skeletal patterns. The evaluated models included an artificial neural network (ANN), Light Gradient Boosting Machine (LightGBM), random forest (RF), extreme gradient boosting (XGBoost), gradient boosting classifier (GBC), decision tree (DT), logistic regression (LR), extra-trees classifier (ET), linear discriminant analysis (LDA), and a support vector machine with a linear kernel (linear SVM). These algorithms were selected to represent different modeling approaches, including linear classifiers, tree-based models, ensemble methods, kernel-based methods, and neural networks. All analyses were conducted using a predefined, patient-level validation pipeline. Patients were randomly divided into training and independent test sets at a ratio of $80\%{:}\,20\%$, respectively. Stratified sampling was used according to the target classification labels to preserve the class distribution in both subsets. When multiple radiographs were available for the same patient, all radiographs from that patient were assigned to the same subset. The independent test set was kept completely separate and was not used for data preprocessing, missing-value imputation, normalization, landmark-coordinate scaling, feature selection, SMOTE resampling, hyperparameter tuning, model selection, or threshold selection. All data-dependent preprocessing steps were incorporated into the modeling pipeline. During cross-validation, each preprocessing step was fitted exclusively on the training portion of each fold and then applied unchanged to the corresponding validation portion. This procedure was used to minimize the risk of information leakage. Hyperparameters were optimized using grid search with stratified five-fold cross-validation performed exclusively within the training set. The search spaces, tested values, and final selected configurations for each model are provided in Table 3. The optimal configuration was selected according to the highest mean validation macro-F1 score. If several configurations produced similar scores, the simpler model was selected according to the predefined tie-breaking rule. A range of supervised machine-learning algorithms was implemented, including an artificial neural network (ANN), LightGBM, random forest (RF), XGBoost, gradient boosting classifier (GBC), decision tree (DT), logistic regression (LR), and extra-trees classifier (ET). These algorithms represent diverse modeling approaches, including linear classifiers, tree-based models, ensemble methods, and neural networks. All analyses were conducted using a patient-level validation pipeline; radiographs from the same patient were assigned to the same subset to ensure independence. Hyperparameters were optimized using grid search with stratified five-fold cross-validation performed exclusively within the training set. The optimal configuration was selected based on the highest mean validation macro-F1 score, with simpler models preferred in cases of similar performance. The ANN was implemented as a feed-forward network (architecture: 64-32 neurons) using ReLU activation in hidden layers and softmax for the output layer. The network was trained using the Adam optimizer (learning rate: 0.001), with L2 regularization (0.01), dropout (0.3), and early stopping (patience: 10) to mitigate overfitting. After tuning, the final models were retrained on the complete training set, and performance was evaluated once on the independent test set.

Leakage-free Analysis

To evaluate the extent to which model performance depended on predictors directly involved in target-label definition, an additional leakage-free analysis was conducted. For sagittal classification, ANB, Wits appraisal, SNA, and SNB were excluded from the predictor set. For vertical classification, the four cephalometric variables directly used to define the vertical skeletal categories (SN–GoGn angle, SN–GoMe angle, Jarabak index [S–Go/N–Me ratio], and sum of posterior angles) were excluded from the predictor set to minimize target leakage. Correlation-based feature selection, imputation, scaling, and SMOTE resampling were strictly restricted to the training data and performed within the relevant training folds. Hyperparameter selection was conducted without access to the independent test set, which was reserved exclusively for final performance evaluation. Because landmark coordinates may still encode morphological information associated with the labeling criteria, this analysis reduces direct label-feature overlap but does not establish complete statistical independence between predictors and target labels.

Table 3. Complete hyperparameter search spaces, cross-validation strategy, and selected configurations for the machine learning models

Model	Hyperparameter	Values evaluated	Selected value
ANN	Hidden-layer architecture	{32}, {64}, {32-64}, {64-32}, {128-64-32}	{128-64-32}
ANN	Learning rate	{0.001, 0.01, 0.1}	0.001
ANN	Batch size	{16, 32, 64, 128}	32
ANN	Dropout rate	{0.0, 0.2, 0.3, 0.5}	0.2
ANN	L2 regularization	{0.0, 0.001, 0.01, 0.1}	0.001
ANN	Maximum epochs	{50, 100, 200}	200
ANN	Early-stopping patience	{5, 10, 15,20}	20
LightGBM	Model-specific parameters	num_leaves: {31, 63}, learning_rate: {0.01, 0.1}, n_estimators: {100, 200}	num_leaves: 31, learning_rate: 0.1, n_estimators: 200
Random Forest	Model-specific parameters	n_estimators: {100, 200, 500}, max_depth: {None, 10, 20}, min_samples_split: {2, 5}	n_estimators: 200, max_depth: 20, min_samples_split: 2
XGBoost	Model-specific parameters	max_depth: {3, 6, 10}, learning_rate: {0.01, 0.1}, n_estimators: {100, 200}	max_depth: 6, learning_rate: 0.1, n_estimators: 200

Table 1 presents the predictor sets used in the primary and leakage-free analyses, distinguishing variables retained as predictors from those excluded to prevent target leakage. For the leakage-free models, landmark coordinates were retained as image-derived anatomical descriptors, whereas cephalometric indices directly used to establish the target labels were excluded.

Results

Demographic Distribution of the Data

A total of 1115 lateral cephalometric radiographs from patients aged 6–18 years were analyzed, including 434 males (38.9%) and 681 females (61.1%). The mean age of male participants was 12.66 ± 3.17years, and that of females was 12.68 ± 3.21years. The age distribution is presented in Figure 3, and the age distribution according to gender is shown in Table 4.

Sagittal and Vertical Labeling: Gender Distribution

Following landmark identification and labeling based on cephalometric indices, the gender distribution within each sagittal and vertical classification is shown in Table 5.

Figure 3. Frequency distribution of each age group among the samples.

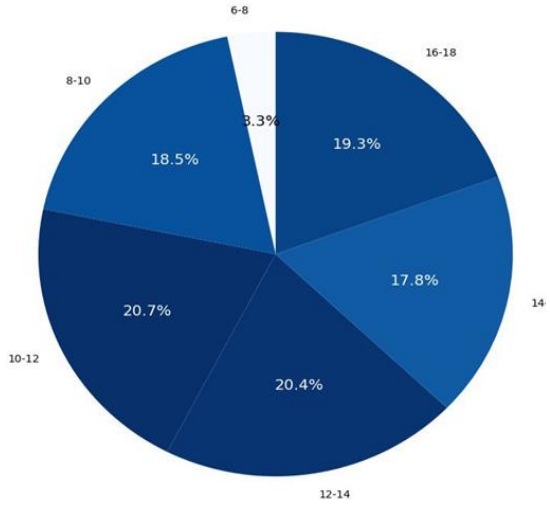

Table 4. Distribution of samples by age and gender

Age (years)	6	7	8	9	10	11	12	13	14	15	16	17	18	Total
Males	–	11	37	37	38	50	49	43	30	33	37	37	32	434
Females	2	21	55	63	65	62	72	48	62	60	71	56	44	681
Total	2	32	92	100	103	112	121	91	92	93	108	93	76	1115

Combined Distribution of Sagittal and Vertical Jaw Relationships

Table 6 shows the combined distribution of sagittal and vertical classes. Across all three sagittal classes, the highest frequency was observed in association with a normal vertical pattern. The long face pattern was most prevalent in sagittal Class III (accounting for 59.2% of all long face cases), while the short face pattern was most frequently observed in sagittal Class I (representing 47.7% of all short face cases).

Intra- and inter-examiner reliability for most variables and indices was > 0.80, supporting the stability of the landmarking process and the validity of the measurements.

Normalization of Coordinates and Exploratory Data Analysis (EDA)

The cephalometric point coordinates (X, Y) were normalized to a 0–1range to eliminate the effect of image scaling. After normalization, scatter maps of points for each class were plotted, and mean and distribution charts were generated for sagittal and vertical classifications (Figures 4 to 8).

A correlation heatmap was also generated for all numerical indices (Figure 9). Strong positive correlations (closer to $+1$) are displayed in darker red, such as the relationship between facial angle and SN–Pog. Conversely, strong negative correlations (closer to -1) are shown in darker blue, such as the correlation between the Jarabak Index and the SN–GoGn angle. Table 7 summarizes pairs of variables with strong correlations ($| r | \geq 0.80$).

Table 5. Sex distribution of data across sagittal and vertical classes

Sagittal class	**Male**	**Female**	**Total**
Class I	167 (36.3%)	293 (63.7%)	460
Class II	143 (40.6%)	209 (59.4%)	352
Class III	124 (40.9%)	179 (59.1%)	303
Total	434 (38.9%)	681 (61.1%)	1115
Vertical class	**Male**	**Female**	**Total**
Normal	295 (37.9%)	483 (62.1%)	778
Long face	29 (40.8%)	42 (59.2%)	71
Short face	110 (41.4%)	156 (58.6%)	266
Total	434 (38.9%)	681 (61.1%)	1115

Table 6. Combined distribution of sagittal and vertical jaw classifications

		Vertical			**Total**
		Normal	Long face	Short face	
Sagittal	Class I	316	17	127	460
	Class II	258	12	82	352
	Class III	204	42	57	303
Total		778	71	266	1115

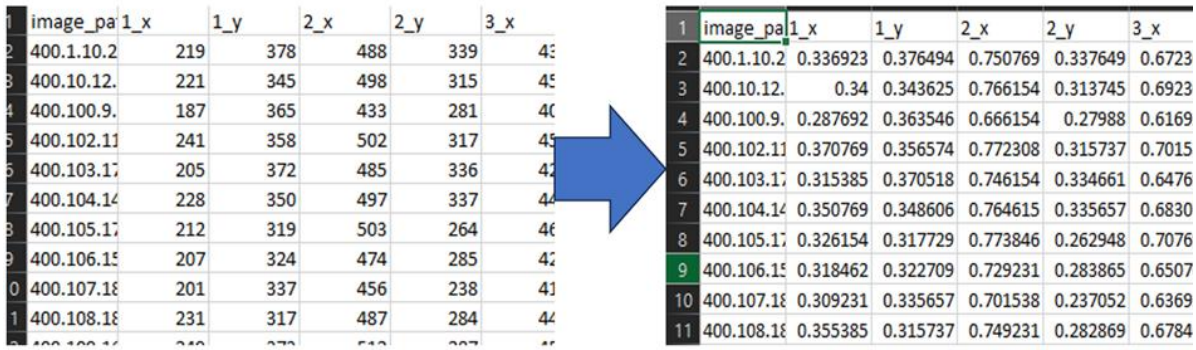

Figure 4. Normalization of selected cephalometric landmark coordinates to a 0–1 range to account for differences in image scale.

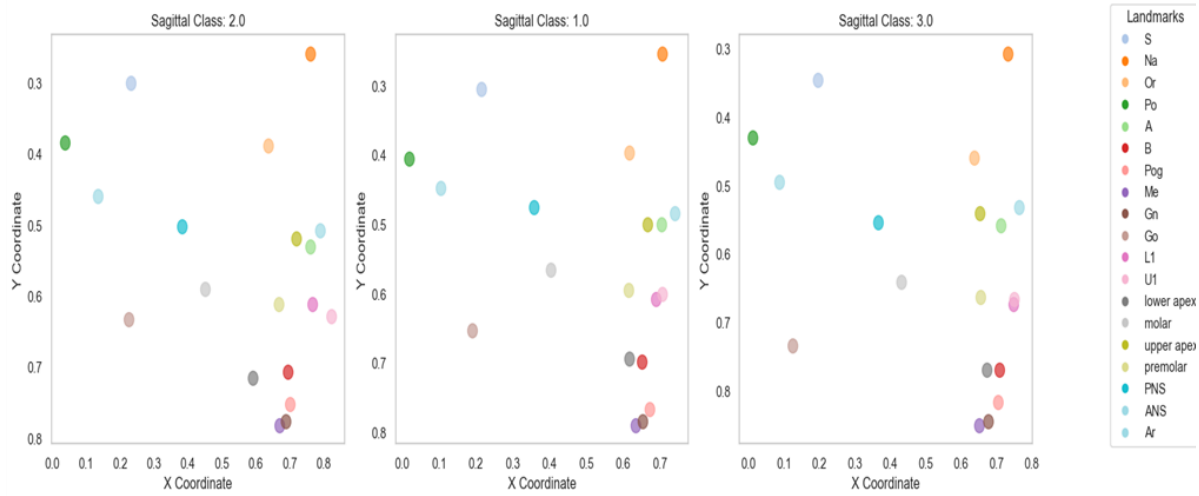

Figure 5. Normalized cephalometric landmark coordinates in representative sagittal Class I, II, and III samples, illustrating the spatial distribution of selected landmarks across skeletal patterns.

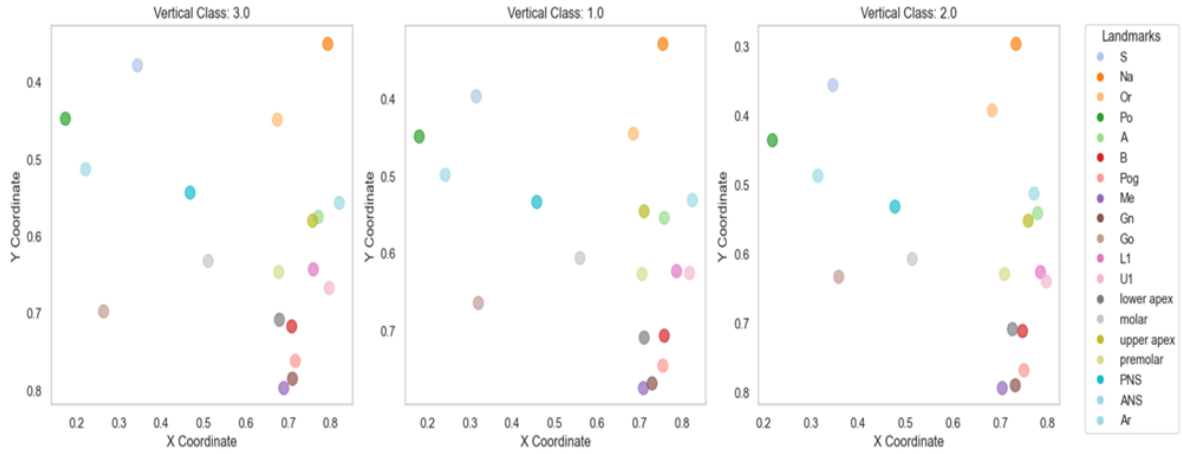

Figure 6. Normalized cephalometric landmark coordinates in representative normal, long face, and short face growth patterns, illustrating differences in vertical skeletal morphology.

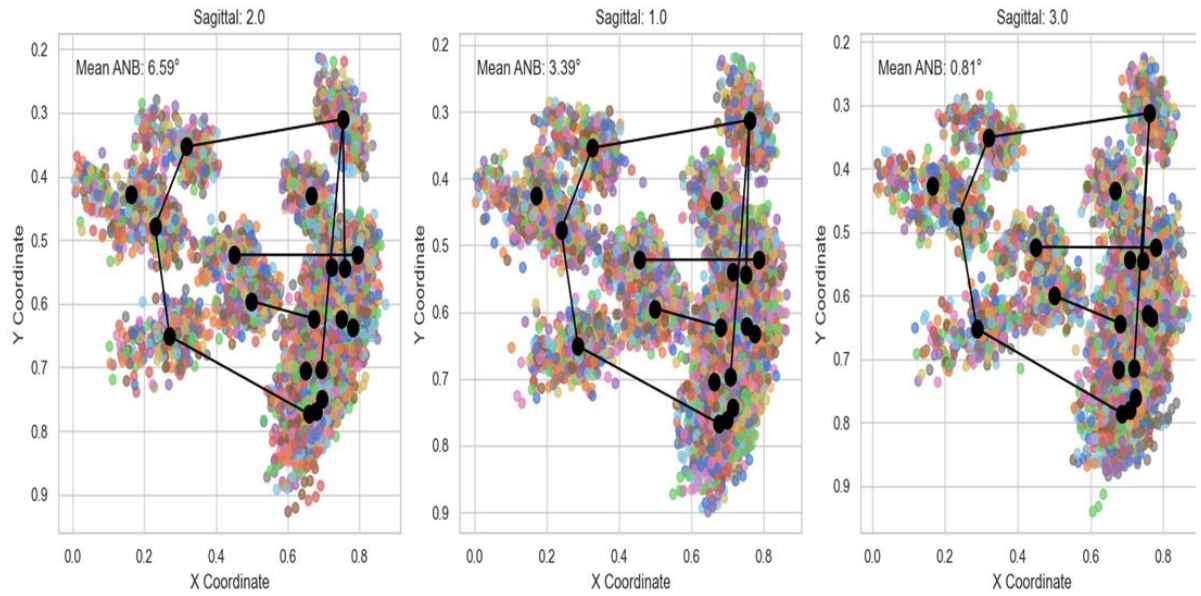

Figure 7. Distribution of normalized landmark coordinates across sagittal classes (Class I, II, and III). Mean ANB angles for each class were 3.39° (Class I), 6.59° (Class II), and 0.81° (Class III).

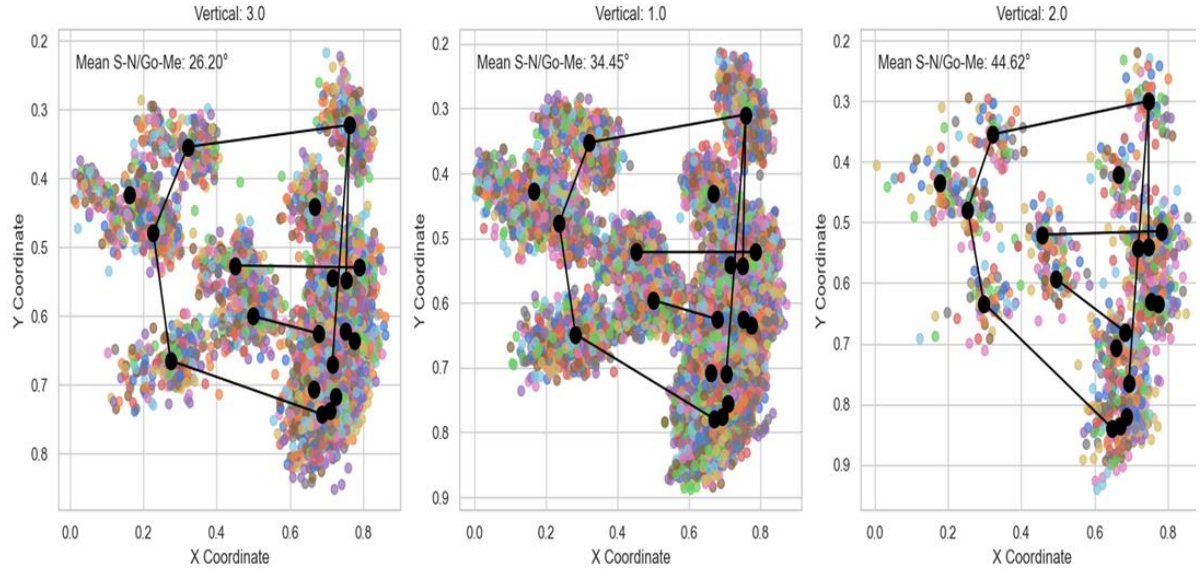

Figure 8. Distribution of normalized landmark coordinates across vertical growth patterns. The mean mandibular plane angle (SN–GoMe) was 34.45° (normal), 44.62° (long face), and 26.2° (short face).

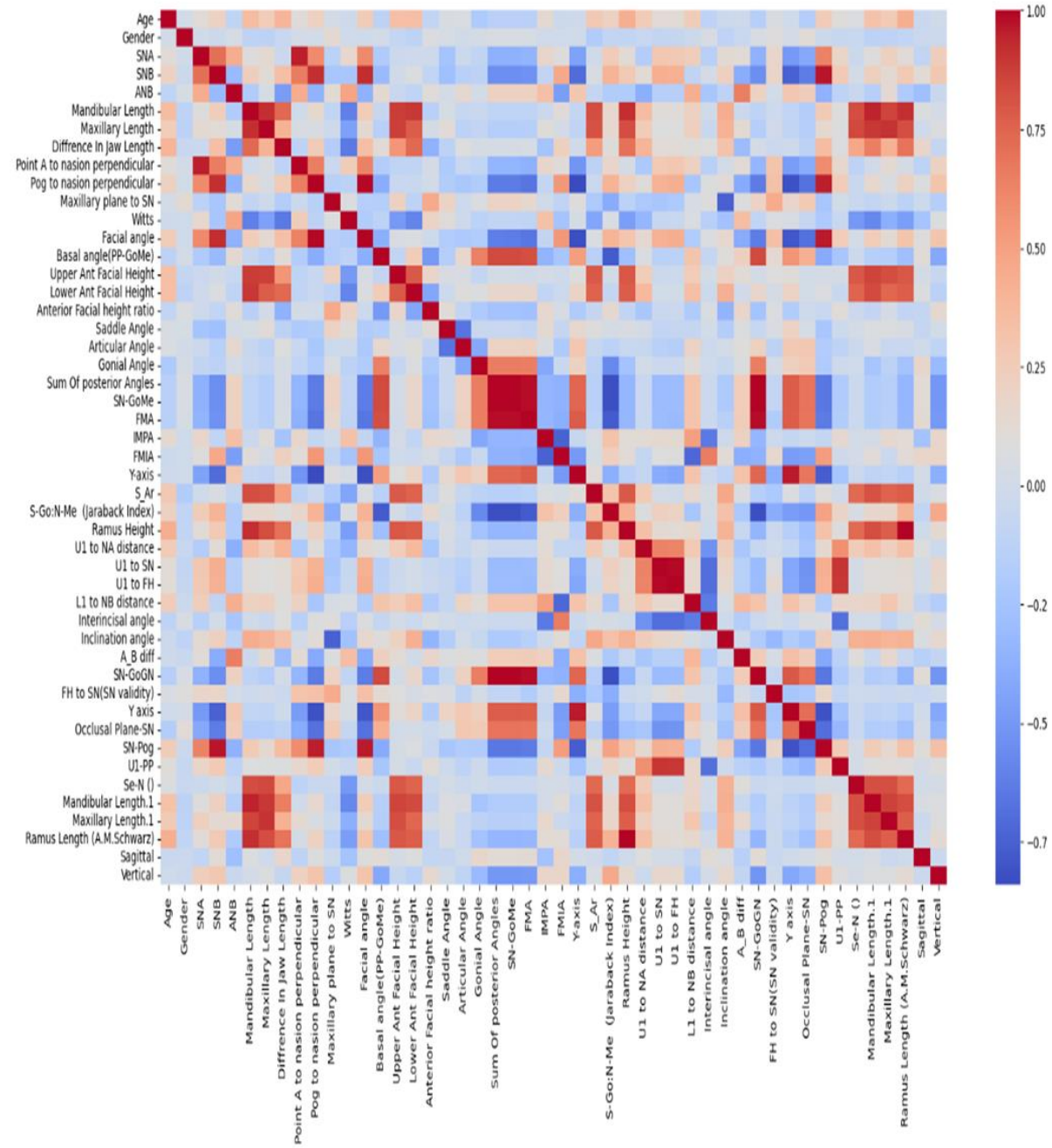

Figure 9. EDA heatmap for linear measurements: each cell illustrates the degree of linear correlation, with stronger positive correlations shown in red and stronger negative correlations in blue.

Table 7. Cephalometric indices with strong correlations (|r| ≥ 0.80). Values of 1.00 indicate near-perfect correlations between redundant or highly dependent measurements (e.g., sum of posterior angles vs. SN–GoMe; ramus height vs. ramus length).

Variable 1	Variable 2	Correlation Coefficient
SNA	Point A to nasion perpendicular	0.95
SNB	Pog to nasion perpendicular	0.92
SNB	Facial angle	0.93
SNB	SN-Pog	0.97
Mandibular length	Maxillary length	0.92
Mandibular length	Upper ant facial height	0.88
Mandibular length	Lower ant facial height	0.89
Mandibular length	S_Ar	0.83
Mandibular length	Ramus height	0.92
Mandibular length	Se-N	0.83
Mandibular length	Mandibular length	0.94
Mandibular length	Maxillary length	0.86
Mandibular length	Ramus length (Co-Go)	0.91
Maxillary length	Upper ant facial height	0.87
Maxillary length	S_Ar	0.82
Maxillary length	Ramus height	0.83
Maxillary length	Se-N	0.85
Maxillary length	Mandibular length	0.89
Maxillary length	Maxillary length	0.91
Maxillary length	Ramus length (Co-Go)	0.83
Pog to nasion perpendicular	Facial angle	0.99
Pog to nasion perpendicular	Y-axis	-0.84
Pog to nasion perpendicular	Y axis	-0.81
Pog to nasion perpendicular	SN-Pog	0.95
Facial angle	Y-axis	-0.82
Facial angle	SN-Pog	0.96
Basal angle (PP-GoMe)	Sum of posterior angles	0.84
Basal angle (PP-GoMe)	SN-GoMe	0.84
Basal angle (PP-GoMe)	FMA	0.82
Basal angle (PP-GoMe)	SN-GoGN	0.84
Upper ant facial height	Mandibular length	0.86
Upper ant facial height	Maxillary length	0.83
Lower ant facial height	Mandibular length	0.84
Sum of posterior angles	SN-GoMe	1.00
Sum of posterior angles	FMA	0.98
Sum of posterior angles	S-Go:N-Me (Jaraback Index)	-0.82
Sum of posterior angles	SN-GoGN	0.99
SN-GoMe	FMA	0.98
SN-GoMe	S-Go:N-Me (Jaraback Index)	-0.82
SN-GoMe	SN-GoGN	0.99
FMA	SN-GoGN	0.98
S_Ar	Mandibular length	0.82
S-Go:N-Me (Jaraback Index)	SN-GoGN	-0.83
Ramus height	Mandibular length	0.83
Ramus height	Ramus length (Co-Go)	1.00
U1 to SN	U1 to FH	0.99
U1 to SN	U1-PP	0.90
U1 to FH	U1-PP	0.89
Y-axis	SN-Pog	-0.81
Se-N	Mandibular length	0.82
Se-N	Maxillary length	0.82
Mandibular length 1	Maxillary length	0.87
Mandibular length 1	Ramus length (Co-Go)	0.83

Machine Learning Model Performance

All machine learning models—including ANN, LightGBM, XGBoost, Random Forest, Gradient Boosting, Decision Tree, Logistic Regression, Extra Trees, and SVM—were trained and evaluated on both sagittal and vertical datasets. The performance of each model was assessed using the independent test data.

Performance of Sagittal Models

In the anteroposterior dimension, LightGBM achieved the best performance, with an accuracy of 0.8510, precision of 0.8570, recall of 0.8510, and an F1 score of 0.8492. The Random Forest and XGBoost models performed comparably, with accuracies around 0.8448. The Decision Tree and Logistic Regression models performed worse than the boosting algorithms, while the Extra Trees model yielded the lowest accuracy (0.7639).

Performance of Vertical Models

In the primary analysis, the ANN achieved 98.96%accuracy for vertical classification. However, some predictors were directly related to the variables used to define the target classes, creating a risk of label leakage. Therefore, this result should be interpreted as performance within the original feature-label configuration rather than as an independent estimate of predictive ability. The leakage-free analysis, which excluded the label-defining measurements, provided a more appropriate assessment of performance beyond these explicit criteria.

The training and validation accuracy and loss curves demonstrated progressive model convergence during ANN development, with no sustained divergence between training and validation performance. The corresponding learning curves are presented in Figure 2.

Overall, the vertical classification task showed higher test performance than the sagittal classification task, with the ANN achieving the highest accuracy in the primary vertical classification analysis.

Table 6. Evaluation of machine learning models for sagittal classification based on accuracy, precision, recall, and F1 score

	Model	Accuracy	Recall	Prec.	F1
LightGBM	Light Gradient Boosting Machine	0.8510	0.8510	0.8570	0.8492
RF	Random Forest Classifier	0.8448	0.8448	0.8536	0.8403
XGBoost	Extreme Gradient Boosting	0.8448	0.8448	0.8522	0.8433
GBC	Gradient Boosting Classifier	0.8447	0.8447	0.8476	0.8427
ANN	Artificial Neural Network	0.8402	0.8402	0.8528	0.8410
DT	Decision Tree Classifier	0.8137	0.8137	0.8204	0.8117
LR	Logistic Regression	0.7989	0.7989	0.8001	0.7979
ET	Extra Trees Classifier	0.7639	0.7639	0.7839	0.7590

Table 7. Evaluation of machine learning models for vertical classification based on accuracy, precision, recall, and F1 score

	Model	Accuracy	Recall	Prec.	F1
ANN	Artificial Neural Network	0.9896	0.9896	0.9903	0.9898
LR	Logistic Regression	0.9121	0.9121	0.9119	0.9092
LightGBM	Light Gradient Boosting Machine	0.9095	0.9095	0.9075	0.9047
XGBoost	Extreme Gradient Boosting	0.9069	0.9069	0.9041	0.9038
GBC	Gradient Boosting Classifier	0.9029	0.9029	0.8987	0.8998
RF	Random Forest Classifier	0.8848	0.8848	0.8820	0.8771
LDA	Linear Discriminant Analysis	0.8822	0.8822	0.8823	0.8792
ET	Extra Trees Classifier	0.8771	0.8771	0.8744	0.8652
SVM	SVM - Linear Kernel	0.8693	0.8693	0.8689	0.8635

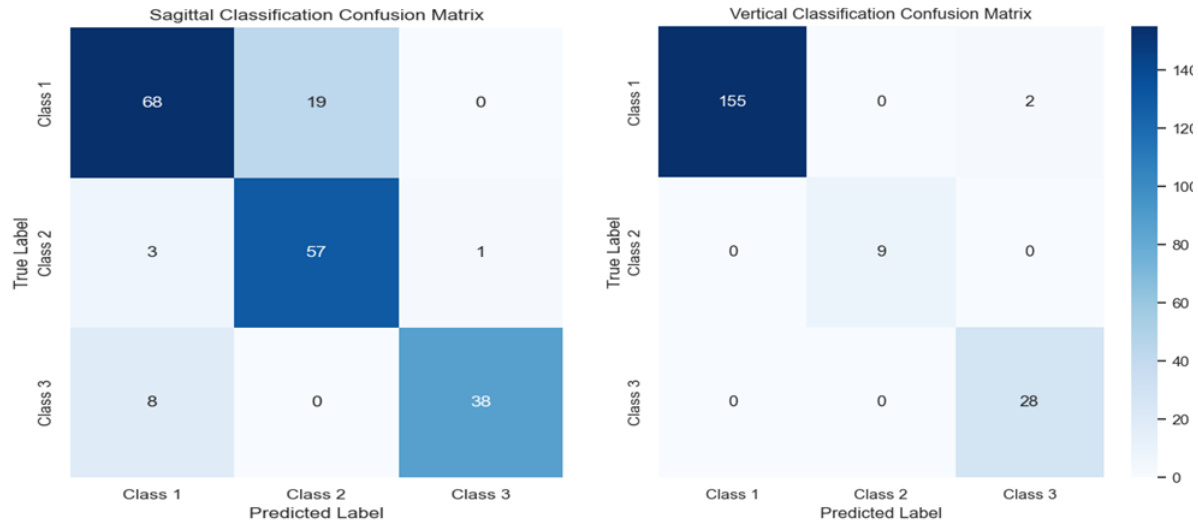

Figure 10. Confusion matrices of test data for sagittal and vertical classification using the ANN model.

Important Predictive Features in Selected Models

Sagittal ANN Model

Permutation-based feature-importance analysis was used to assess the relative contribution of input variables to model predictions. For sagittal classification, the ANN model identified ANB, the X-coordinates of points A and B, the X-coordinate of the mandibular incisor apex, Wits appraisal, the X-coordinates of Me, Pog, and N, and SNA and SNB as the most influential predictors (Figure 11).

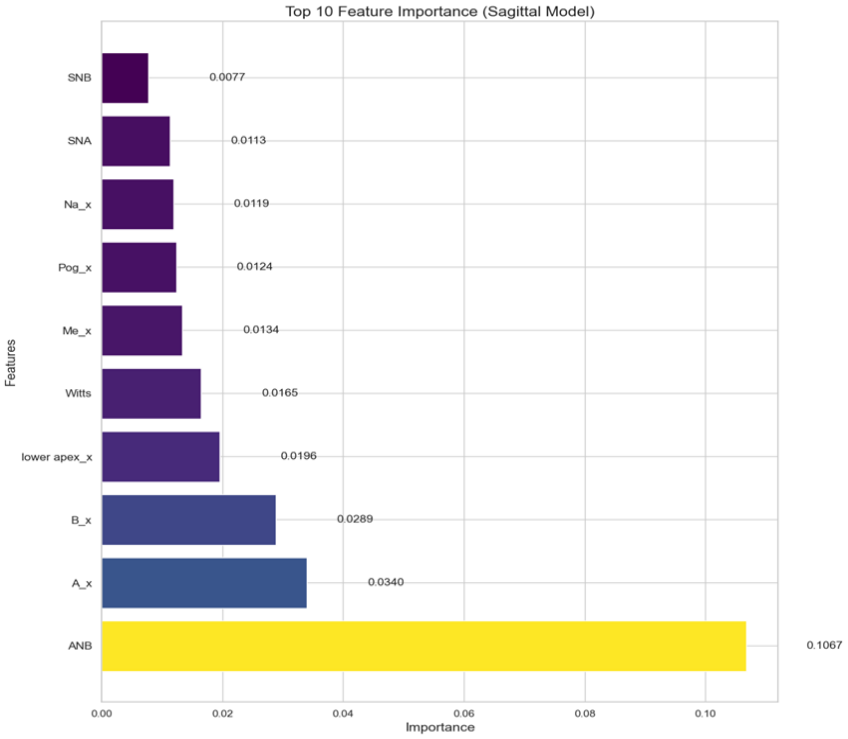

Figure 11. Permutation-based feature importance of the ANN model for sagittal skeletal pattern classification.

Vertical ANN Model

For vertical classification, age, the Y-coordinates of Gn and Pog, the Y-coordinate of the maxillary incisor apex, the X-coordinate of the molar contact point, and the sum of posterior angles were among the most influential predictors (Figure 12).

Sagittal LightGBM Model

For the sagittal LightGBM model, the most influential predictors, in descending order, were ANB, Wits appraisal, sum of posterior angles, age, SNA, SNB, and the basal, saddle, and gonial angles (Figure 13).

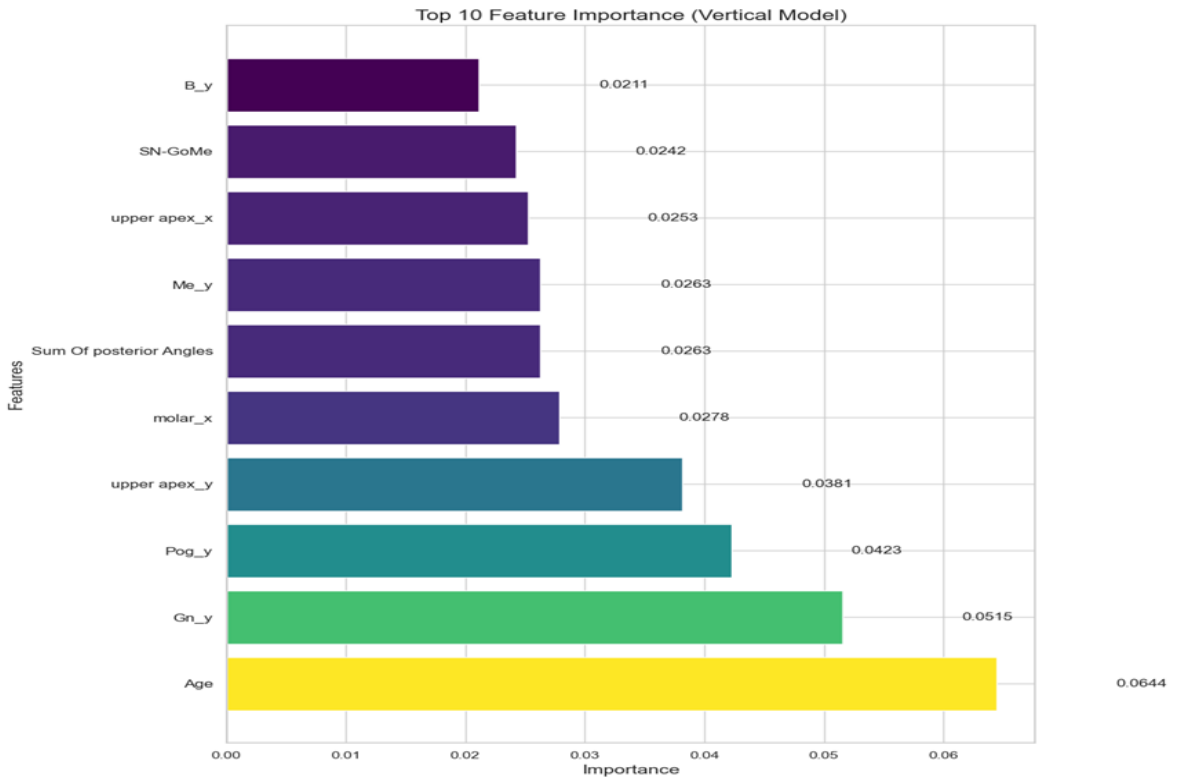

Figure 12. Permutation-based feature importance of the ANN model for vertical skeletal pattern classification.

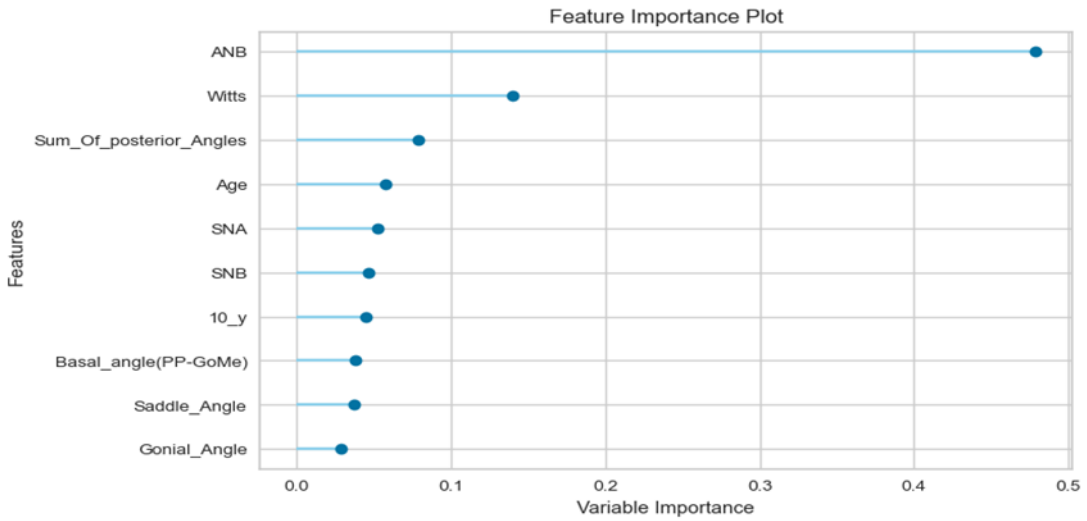

Figure 13. Important features identified by the sagittal LightGBM model.

Vertical Logistic Regression Model

In the vertical logistic regression model, age was the most influential predictor, followed by the Y-coordinates of Gn and Pog and the Y-coordinate of the maxillary incisor apex (Figure 14).

Imbalance-aware Evaluation and Class-wise Performance

Because the skeletal classes were unequally distributed, class-wise and imbalance-aware metrics were calculated for both classification tasks. SMOTE was applied only to the training folds; the independent test sets retained their original class distributions.

For sagittal classification, the test set comprised 92 Class I, 70 Class II, and 61 Class III cases (total $n = 223$). The ANN model achieved a macro-F1 score of 85.1%, and a macro ROC-AUC of 0.9184. Class-wise sensitivity values were 91.3%, 80.0%, and 82.0% for Class I, Class II, and Class III, respectively. The corresponding specificity values were 93.1%, 90.8%, and 92.5% respectively. Precision values were 89.7%, 82.8%, and 85.1% while the corresponding F1-scores were 90.5%, 81.4%, and 83.5% respectively. Most classification errors occurred between adjacent sagittal categories, particularly between Class I and Class II.

For vertical classification, the test set consisted of 156 normal, 14 long face, and 53 short face cases (total n = 223). The ANN model achieved a balanced accuracy of 96.2% and a macro-F1 score of 95.0%. Sensitivity was 99.4% for the normal group, 92.9% for the long face group, and 96.2% for the short face group. Corresponding specificity values were 98.5%, 99.0%, and 98.7%, respectively. Although the long face category was the smallest subgroup in the independent test set, the model maintained high class-specific performance after SMOTE was applied exclusively to the training folds. Tables 8 and 9 present the detailed class-wise performance metrics.

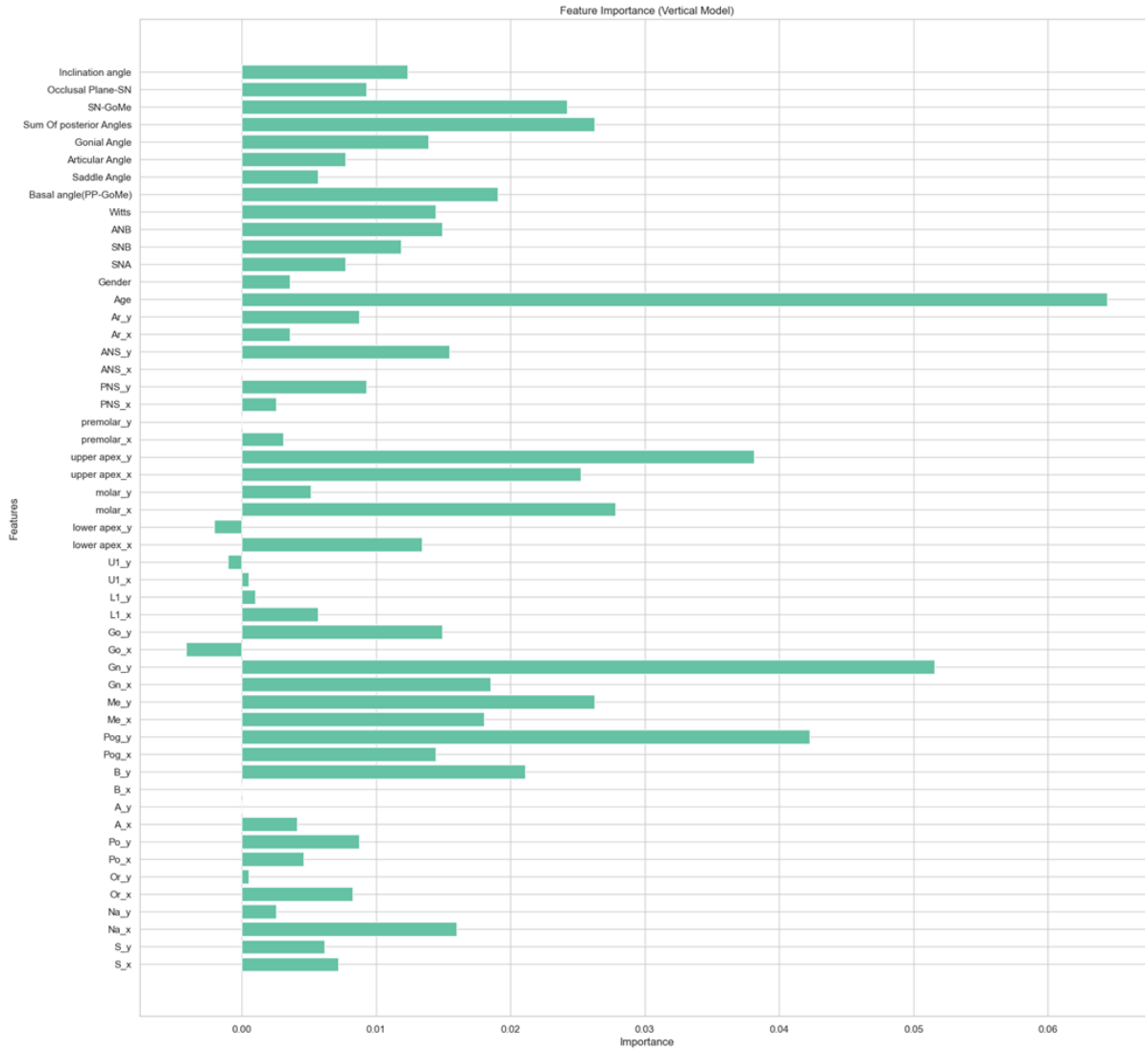

Figure 14. Important features identified by the vertical Logistic Regression model.

Table 8. Performance of machine-learning models for sagittal classification

Class	Test samples	Sensitivity	Specificity	Precision	F1-score	Macro ROC-AUC (OvR)
Class I	92	91.3%	93.1%	89.7%	90.5%	-
Class II	70	80.0%	90.8%	82.8%	81.4%	-
Class III	61	82.0%	92.5%	85.1%	83.5%	-
Macro average	223	—	—	—	85.1%	0.9184

Table 9. Performance of machine-learning models for vertical classification

Class	Test samples	Sensitivity	Specificity	Precision	F1-score	Macro ROC-AUC (OvR)
Normal	156	99.4%	98.5%	98.7%	99.0%	-
Long-face	14	92.9%	99.0%	86.7%	89.7%	-
Short-face	53	96.2%	98.7%	96.2%	96.2%	-
Macro average	223	—	—	—	95.0%	0.9987

Leakage-free Analysis

The primary analysis used the full predictor set after feature selection based on the training data. In the leakage-free analysis, cephalometric variables directly used to define the target labels were excluded. For sagittal classification, ANB, Wits appraisal, SNA, and SNB were excluded. For vertical classification, SN–GoGn, SN–GoMe, the Jarabak index, and the sum of posterior angles were excluded. Landmark coordinates were retained in both analyses as image-derived anatomical descriptors.

Compared with the primary analysis, performance was lower in the leakage-free analysis. For sagittal classification, LightGBM achieved the best reported performance, with an accuracy of 78.5%, macro-precision of 78.9%, macro-recall of 78.3%, and macro-F1 score of 78.5%. For vertical classification, ANN achieved an accuracy of 82.5%, macro-precision of 82.0%, macro-recall of 82.7%, and macro-F1 score of 82.0%.

The reduction was greater for vertical classification when compared with the primary vertical ANN accuracy of 98.96%. This suggests that the primary vertical performance was partly influenced by predictors related to the label definitions. Accordingly, the primary analysis should be interpreted as a pattern-recognition and consistency-support assessment rather than as an independent diagnostic evaluation. Class-wise results for the best-performing leakage-free models are shown in Table 10.

Table 10. Class-wise performance of the best-performing leakage-free models for sagittal and vertical skeletal classification

Task	Model	Class	Test samples	Sensitivity	Specificity	Precision	F1-score
Sagittal	LightGBM	Class I	92	81.5%	82.5%	76.5%	79.0%
Sagittal	LightGBM	Class II	70	71.4%	92.8%	82.0%	76.4%
Sagittal	LightGBM	Class III	61	82.0%	91.4%	78.1%	80.0%
Sagittal	LightGBM	Macro-average	223	78.3%	—	78.9%	78.5%
Sagittal	LightGBM	Accuracy	223	—	—	—	78.5%
Vertical	ANN	Normal	156	83.3%	80.6%	90.9%	87.0%
Vertical	ANN	Long face	14	85.7%	99.5%	92.3%	88.9%
Vertical	ANN	Short face	53	79.2%	85.3%	62.7%	70.0%
Vertical	ANN	Macro-average	223	82.7%	—	82.0%	82.0%
Vertical	ANN	Accuracy	223	—	—	—	82.5%

Discussion

This study evaluated 1,115 lateral cephalograms from 6–18-year-old Iranian patients to assess the performance of machine learning (ML) models in classifying sagittal (Class I, II, and III) and vertical (normal, long face, and short face) skeletal relationships. LightGBM achieved the highest performance in sagittal classification, whereas the ANN performed best in vertical classification. This difference may reflect variations in model architecture and the distribution of input features across the two tasks. The findings support the potential of ML as a structured, semi-automated framework for skeletal pattern recognition in growing patients, rather than as a replacement for conventional cephalometric assessment (17,18). However, the performance estimates, particularly for vertical classification, should be interpreted in light of the overlap between some input features and the variables used to define the target labels.

The most influential features in the primary sagittal models included the ANB angle, Wits appraisal, and coordinates of points A and B, whereas age, the Y-coordinates of Gn and Pog, and the sum of posterior angles were prominent in vertical classification. These findings suggest that Machine learning models can capture patterns associated with skeletal relationships. Nevertheless, feature importance should be interpreted cautiously because correlated predictors may share or substitute for predictive

information, making importance rankings dependent on the model and dataset. Permutation importance, therefore, reflects model-specific predictive contribution rather than biological hierarchy, causal relationships, or clinical significance.

SMOTE was applied only to the training data to address class imbalance and avoid contaminating the independent test set. Nevertheless, resampling may influence learned decision boundaries, particularly for the minority vertical class. Accordingly, overall accuracy should be interpreted alongside class-specific performance, and results for less represented categories warrant particular caution.

The primary vertical ANN achieved 98.96% accuracy; however, this result should not be interpreted as an independent estimate of predictive performance because some predictors overlapped with the variables used to define the target labels. In the leakage-free analysis, the ANN achieved 82.5% accuracy and a macro-F1 score of 82%. The lower performance suggests the primary estimate was optimistic and partly reflected overlap between predictors and labeling criteria. Accordingly, the leakage-free results provide a more appropriate estimate of internal predictive performance, although they do not establish clinical validity or generalizability. Further evaluation using landmark-coordinate-only models and independent external datasets is needed.

The use of a relatively large, anonymized dataset from a single institution provides evidence from an Iranian cohort and highlights the importance of considering population-specific variation in cephalometric norms (15). The high intra- and inter-examiner reliability (ICC>0.8) supports the consistency of the measurements used for model development. Nevertheless, measurement reliability may vary with image quality and examiner experience, emphasizing the importance of ongoing quality assurance.

Comparisons with previous studies provide context for the observed performance, although differences in populations, target definitions, predictor sets, and validation strategies limit direct comparisons of accuracy. For sagittal classification, the LightGBM accuracy of 85.1% was lower than the 90.3% reported for DenseNet-121 in children using lateral cephalograms and the approximately 89% reported for ResNet-152 in adults (17,19). Differences in sample composition and classification criteria may partly explain this variation. In particular, borderline cases may be harder to classify; a previous study reported an increase to 94% after excluding ambiguous cases (17). However, excluding such cases may reduce a model's applicability to routine clinical populations, where borderline skeletal relationships are also encountered.

The interpretation of sagittal classification depends partly on the limitations of the indices used to establish the reference labels. ANB and Wits appraisal were used to classify sagittal relationships, with Wits serving as a complementary measure when ANB was considered less reliable, including cases involving abnormal SNA and SNB values or long face patterns (20). ANB may be affected by cranial base length, mandibular rotation, and vertical facial pattern, whereas Wits is sensitive to occlusal plane inclination and dental compensation and should not be used in isolation (21). Combining these indices with landmark coordinates provides a broader representation of sagittal relationships. However, because some labeling-related variables were also included as predictors in the primary models, the importance attributed to these features should not be interpreted as evidence of independent diagnostic value. In the leakage-free analysis, landmark coordinates and demographic variables provided alternative sources of information for predicting the predefined skeletal categories without directly including the label-defining measurements. However, these predictors may still encode morphological information associated with the reference labels.

Previous studies have also demonstrated the potential of ML for cephalometric classification. In a German cohort with skeletal Class I and III patterns, Random Forest and CART models achieved high performance using multiple cephalometric parameters, while reduced predictor sets also yielded favorable results (22).

Machine learning models have similarly shown high classification performance in Arab patients with skeletal Class I and III patterns (23). For vertical assessment, a hybrid deep-learning model combining EfficientNet-B0 and DenseNet201 achieved accuracies of 86.67% for FMA, 87.29% for SN-GoGn, and 82.71% for the occlusal plane, supporting the feasibility of direct classification from lateral cephalograms without explicit landmark detection (24). Another German study reported high performance in classifying skeletal Classes I, II, and III using Calculated_ANB as the reference classification (25). Collectively, these findings support the feasibility of ML-based cephalometric pattern recognition while emphasizing the importance of considering reference labels and predictor definitions when interpreting performance across studies. In sagittal

classification, the prominence of ANB and Wits in the primary models is also consistent with previous ML research on skeletal Class III patterns, in which cephalometric variables related to sagittal relationships showed substantial predictive value (13).

The observed relationships between cephalometric variables are also relevant to orthodontic interpretation. The strong correlations among SNB, facial angle, and SN-Pog indicate substantial overlap among several sagittal measurements, whereas the negative correlation between the Jarabak Index and SN-GoGn is consistent with the relationship between vertical facial morphology and mandibular-plane inclination. The prominent contribution of age in the vertical ANN may indicate that growth-related changes provide information beyond individual static cephalometric measurements (15). However, growth-stage stratification was not performed, so the independent contribution of developmental stage could not be fully assessed. These findings should therefore be interpreted as associations within the evaluated dataset rather than evidence of specific biological mechanisms.

Limitations

Several limitations should be considered. First, the study relied on two-dimensional cephalometric radiographs, in which anatomical superimposition may affect landmark localization (26). Second, manual landmark identification remains susceptible to measurement error, which may propagate through automated workflows without mechanisms for real-time correction (18). Third, the single-center design and lack of external validation limit how far the findings can be generalized to other populations. Potential multicollinearity may also affect model stability, despite feature selection. In addition, the cross-sectional design does not permit assessment of longitudinal craniofacial growth. Finally, partial overlap between input features and labeling criteria in the primary analysis may have produced optimistic performance estimates, reinforcing the need for leakage-free evaluation.

Future research should prioritize leakage-free models, including approaches based on landmark coordinates and demographic variables, and evaluate them using independent, multicenter datasets representing diverse populations. Longitudinal datasets and age-adjusted or growth-aware classifiers may help clarify the contribution of developmental changes to skeletal classification. Multimodal approaches incorporating complementary information, such as profile photographs or three-dimensional facial data when clinically justified, may also improve model robustness. Transparent reporting of class-specific performance and external validation will be essential before potential clinical applications can be established.

Overall, this study provides a proof of concept for ML-assisted cephalometric pattern recognition in growing patients. Although the primary analyses showed promising performance, partial overlap between predictors and label-defining variables limits the interpretation of those results. The leakage-free analyses provide a more appropriate estimate of predictive performance; however, independent validation is still required before the models can be considered generalizable or suitable for routine clinical use.

Conclusion

This study evaluated machine-learning models for classifying sagittal and vertical skeletal patterns in growing patients using lateral cephalometric data. LightGBM showed the strongest performance for sagittal classification, whereas the ANN performed best for vertical classification in the primary analysis. However, the high vertical-classification performance may have been influenced by the inclusion of predictors related to the measurements used to define the target categories. After excluding the explicitly label-defining cephalometric indices, ANN performance decreased to 82.5% accuracy and a macro-F1 score of 82.0%. These findings indicate that machine learning may assist with structured cephalometric pattern recognition, but they do not demonstrate clinical validity, superiority over conventional assessment, or generalizability beyond the study population. Because landmark coordinates remained in the reduced-leakage analysis, the findings should be interpreted as an initial internal evaluation rather than as evidence of independent diagnostic performance. Further studies using prespecified coordinate-based models, geographically and institutionally diverse datasets, and external validation are needed before the clinical usefulness of these models can be determined. External validation should assess not only discrimination but also calibration and clinical utility before implementation is considered.

Acknowledgments
None

Conflict of Interests
The authors declare no conflict of interest.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Funding

This research received no external funding.

Authors' Contributions

Kazem Dalaie: Conceptualization, Methodology, Writing – Original Draft, Project Administration.

Shahab Kavousinejad: Conceptualization, Methodology, Software Development, Formal Analysis, Statistical Analysis, Writing – Original Draft, Project Administration.

Fateme Aliverdi: Data Collection, Expert Labeling, Data Curation, Methodology, Writing – Original Draft.

Nazila Biglar: Literature Review, Methodology, Writing – Original Draft.

References

1. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769-774. doi: 10.1177/0022034520915714.
2. Subramanian AK, Chen Y, Almalki A, Sivamurthy G, Kafle D. Cephalometric analysis in orthodontics using artificial antelligence-a comprehensive review. Biomed Res Int. 2022;2022:1880113. doi: 10.1155/2022/1880113.
3. Junaid N, Khan N, Ahmed N, Abbasi MS, Das G, Maqsood A, Ahmed AR, etal. Development, application, and performance of artificial intelligence in cephalometric landmark identification and diagnosis: a systematic review. Healthcare (Basel). 2022;10(12):1-14. doi: 10.3390/healthcare10122454.
4. Ueda A, Tussie C, Kim S, Kuwajima Y, Matsumoto S, Kim G, etal. Classification of maxillofacial morphology by artificial intelligence using cephalometric analysis measurements. Diagnostics (Basel). 2023;13(13):1-12. doi: 10.3390/diagnostics13132134.
5. Kunz F, Stellzig-Eisenhauer A, Zeman F, Boldt J. Artificial intelligence in orthodontics : Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. J Orofac Orthop. 2020;81(1):52-68. English. doi: 10.1007/s00056-019-00203-8.
6. Hwang HS, Kim WS, McNamara JA Jr. Ethnic differences in the soft tissue profile of Korean and European-American adults with normal occlusions and well-balanced faces. Angle Orthod. 2002;72(1):72-80. doi: 10.1043/0003-3219(2002)072<0072:EDITST>2.0.CO;2.
7. Balel Y, Sağtaş K, Teke F, Kurt MA. Automated classification of sagittal, vertical, and transverse malocclusions from 3D intraoral scans via a multi-head CNN framework. Odontology. 2026;114(4):2722-2732. doi: 10.1007/s10266-026-01400-9.
8. Dashti M, Khosraviani F, Esmaeili S, Ghadimi N, Orhan K, Meyari A, etal. Artificial intelligence for skeletal classification in orthodontics: A systematic review and meta-analysis of lateral cephalometric studies. J Taibah Univ Med Sci. 2026 ;21(4):683-697. doi: 10.1016/j.jtumed.2026.06.003.
9. Ajayi EO. Cephalometric norms of Nigerian children. Am J Orthod Dentofacial Orthop. 2005;128(5):653-6. doi: 10.1016/j.ajodo.2005.07.002.
10. Guinot-Barona C, Alonso Pérez-Barquero J, Galán López L, Barmak AB, Att W, Kois JC, etal. Cephalometric analysis performance discrepancy between orthodontists and an artificial intelligence model using lateral cephalometric radiographs. J Esthet Restor Dent. 2024;36(4):555-565. doi: 10.1111/jerd.13156.
11. Polizzi A, Leonardi R. Automatic cephalometric landmark identification with artificial intelligence: An umbrella review of systematic reviews. J Dent. 2024;146:105056. doi: 10.1016/j.jdent.2024.105056.
12. Nordblom NF, Büttner M, Schwendicke F. Artificial intelligence in orthodontics: critical review. J Dent Res. 2024;103(6):577-584. doi: 10.1177/00220345241235606.
13. Zhao L, Chen X, Huang J, Mo S, Gu M, Kang N, etal. Machine learning algorithms for the diagnosis of class III malocclusions in children. Children (Basel). 2024;11(7):1-14. doi: 10.3390/children11070762.
14. Surendran A, Daigavane P, Shrivastav S, Kamble R, Sanchla AD, Bharti L, etal. The Future of orthodontics: deep learning technologies. Cureus. 2024;16(6):1-13. doi: 10.7759/cureus.62045.
15. Azarbayejani S, Omrani A, Kalaantar-Motamedi A, Abdellahi M, Taalebi V, etal. Cephalometric norms for 6-17 year-old Iranians with normal occlusion and well-balanced faces. Dent Res J (Isfahan). 2014;11(3):327-35.doi: 10.4103/1735-3327.132023.
16. Jacobson A. The "Wits" appraisal of jaw disharmony. Am J Orthod. 1975;67(2):125-38. doi: 10.1016/0002-9416(75)90065-2.

17. Nan L, Tang M, Liang B, Mo S, Kang N, Song S, etal. Automated sagittal skeletal classification of children based on deep learning. Diagnostics (Basel). 2023;13(10):1-14. doi: 10.3390/diagnostics13101719.
18. Schwendicke F, Chaurasia A, Arsiwala L, Lee JH, Elhennawy K, Jost-Brinkmann PG, etal. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. Clin Oral Investig. 2021;25(7):4299-4309. doi: 10.1007/s00784-021-03990-w.
19. Li H, Xu Y, Lei Y, Wang Q, Gao X. Automatic classification for sagittal craniofacial patterns based on different convolutional neural networks. Diagnostics (Basel). 2022;12(6):1-11. doi: 10.3390/diagnostics12061359.
20. Paddenberg E, Proff P, Kirschneck C. Floating norms for individualising the ANB angle and the WITS appraisal in orthodontic cephalometric analysis based on guiding variables. J Orofac Orthop. 2023;84(1):10-18. doi: 10.1007/s00056-021-00322-1.
21. Boccalari E, Rossi O, Baldini B, Tripicchio C, Serafin M, Caprioglio A. The maxillomandibular sagittal assessment: The ABwise appraisal and its correlation with ANB angle. J Clin Med. 2025;14(4):1-10. doi: 10.3390/jcm14041379.
22. Paddenberg-Schubert E, Midlej K, Krohn S, Schröder A, Awadi O, Masarwa S, etal. Machine learning models for improving the diagnosing efficiency of skeletal class I and III in German orthodontic patients. Sci Rep. 2025;15(1):12738. doi: 10.1038/s41598-025-97717-6.
23. Midlej K, Zohud O, Masarwa S, Paddenberg-Schubert E, Krohn S, Proff P, etal. Applying machine-learning models on lateral cephalogram parameters to classify skeletal classes I and III arab orthodontic patients, instead of individualized parameters. J Int Soc Prev Community Dent. 2025;15(5):474-484. doi: 10.4103/jispcd.jispcd_231_24.
24. Özel MB, Kartbak SBA, Çakmak M. Classification Performance of deep learning models for the assessment of vertical dimension on lateral cephalometric radiographs. Diagnostics (Basel). 2025;15(17):1-16. doi: 10.3390/diagnostics15172240.
25. Paddenberg-Schubert E, Midlej K, Krohn S, Kuchler E, Watted N, Proff P, etal. Automated classification of skeletal malocclusion in German orthodontic patients. Clin Oral Investig. 2025;29(8):1-15. doi: 10.1007/s00784-025-06485-0.
26. Hendrickx J, Gracea RS, Vanheers M, Winderickx N, Preda F, Shujaat S, etal. Can artificial intelligence-driven cephalometric analysis replace manual tracing? A systematic review and meta-analysis. Eur J Orthod. 2024;46(4):1-17. doi: 10.1093/ejo/cjae029.