Zecca *et al. BMC Oral Health* (2025) 25:1524
https://doi.org/10.1186/s12903-025-06925-z

BMC Oral Health

RESEARCH Open Access

An open-source AI tool for predicting cephalometric measurements from clinical data and photographic images

Piero Antonio Zecca[1*], Margherita Caccia[1,2], Luca Levrini[3], Andrea Carganico[4], Marco Serafin[5], Petra Rita Basso[1] and Marcella Reguzzoni[1]

Abstract

Introduction Traditional orthodontic diagnostics rely significantly on lateral cephalometric radiographs, posing health risks due to ionising radiation, particularly in paediatric patients. Artificial intelligence (AI) represents a promising alternative by enabling predictions of cephalometric parameters from non-radiographic clinical data. This study evaluates the accuracy and clinical utility of CEPHCLINIC, an open-source AI software designed to predict conventional cephalometric measurements using clinical photographs and intraoral 3D scans, thus adhering to radiation protection principles.

Materials and methods The dataset comprised 1255 subjects from the American Association of Orthodontists Foundation (AAOF) craniofacial collection, encompassing demographic and clinical variables (age, gender, overbite, overjet, facial dimensions). This dataset was randomly divided into training (80%, $n=1004$) and validation (20%, $n=251$) subsets. Additionally, an independent external test set of 51 untreated orthodontic cases was employed for rigorous evaluation. Input variables for model training included clinical parameters derived from photographs (WebCeph software) and intraoral scans (iTero scanner, MeshMixer software). Supervised predictive regression models, including ExtraTreesRegressor, CatBoostRegressor, and Support Vector Regression, were optimised through GridSearchCV and validated using repeated random subsampling. Predictive accuracy was assessed statistically using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Spearman correlation coefficients, R-squared values, and paired t-tests.

Results The ExtraTreesRegressor demonstrated superior performance across multiple cephalometric parameters, achieving notably low RMSE values in the independent test set for ANB (2.772°) and NP2PA (2.317 mm). However, parameters like COPAD exhibited higher prediction errors (RMSE 12.121 mm). Spearman correlation analysis indicated strong prediction consistency for COPOD (0.850), moderate for U1SNA (0.548), and poor predictability for NP2PO (-0.052). Despite statistically significant biases observed in predictions for some parameters (paired t-test, $p<0.05$), overall predictive accuracy was clinically acceptable, emphasising parameters such as COPOD, COPAD, ANB, and U1SNA as particularly reliable.

*Correspondence:
Piero Antonio Zecca
pieroantonio.zecca@uninsubria.it

Full list of author information is available at the end of the article

Conclusions The CEPHCLINIC software reliably predicts key cephalometric measurements from non-radiographic clinical data, significantly reducing radiation exposure risks. Despite promising performance, model refinement, dataset expansion with broader demographic representation, and integration with digital orthodontic technologies are essential for enhancing precision, clinical reliability, and global applicability.

Keywords Artificial intelligence, Cephalometry, Digital radiology, Open-source software, Prediction

Introduction

Incorporating artificial intelligence (AI) into orthodontic diagnostics represents a significant advancement toward standardised and more efficient cephalometric analyses [1, 2]. Traditionally, orthodontists depend on lateral cephalometric radiographs to assess craniofacial structures accurately and devise appropriate treatment plans. However, reliance on radiographic imaging introduces potential health risks from ionising radiation exposure, particularly significant in paediatric populations [3]. Consequently, adhering to radioprotection principles such as ALADAIP, which advocate minimising radiation exposure without compromising diagnostic accuracy, has become essential nowadays [4].

AI technologies, particularly machine learning and deep learning, may effectively address these radioprotection concerns by enabling accurate cephalometric predictions from non-radiographic clinical data. Specifically, convolutional neural networks have shown promise in automating the detection and analysis of various cephalometric landmarks [5]. This automation not only enhances diagnostic precision but may significantly reduce unnecessary patient radiation exposure, aligning closely with radioprotection principles.

Recent literature highlights the robust capabilities of AI in providing consistent and accurate cephalometric analyses across diverse patient populations, demonstrating its reliability and applicability in orthodontic practice [6]. Zecca et al. previously demonstrated AI-driven approaches could accurately derive cephalometric parameters even from limited-field radiographs, substantially reducing radiation exposure while maintaining diagnostic integrity [7]. These findings underscore the potential of AI to enhance patient safety by substantially decreasing radiation exposure in orthodontic diagnostics.

Additionally, AI contributes significantly to the prediction of orthodontic treatment outcomes [8]. By analysing extensive clinical datasets, AI systems can assist clinicians in developing individualised treatment plans tailored to patients' unique craniofacial characteristics and growth patterns. This individualised approach aligns closely with modern standards of personalised care, minimal invasiveness, and enhanced time-consuming.

The present study aimed to evaluate the reliability of predicting traditional cephalometric parameters without radiographic imaging, utilising AI methodologies applied to clinical photographs and intraoral scans. Specifically, this research introduced the CEPHCLINIC tool, an innovative open-source software designed to adhere to ALADAIP principles by minimising radiation exposure and predicting cephalometric-inspired parameters from routinely collected clinical images.

Materials and methods

The present research received ethical approval from the Ethics Committee of the University of Insubria (protocol number 0026262, issued on 21/02/2025) and adhered to the guidelines set forth by the Helsinki Declaration.

The primary objective of this study was to comparatively assess traditional cephalometric parameters derived from standard lateral cephalometric radiographs against estimations generated through AI-based predictive models. These supervised predictive models were trained using input variables derived exclusively from clinical photographs and 3D dental scans. The corresponding output-target variables were actual cephalometric parameters obtained from traditional lateral cephalometric radiographs, used as ground truth during model training.

Data utilised for training the predictive algorithms were sourced from the craniofacial growth legacy collection of the American Association of Orthodontists Foundation (AAOF). All cephalometric variables measurable from non-radiographic data sources, such as photographs, dental scans, and patient demographics, were included as input features. Specifically, features like overbite (OB), overjet (OJ), upper 6 to lower 6 distance (U6L6D), lower face height (LFHNP), interincisal angle (IIANG), total face height (TFHNP), and percent lower face height (PLHNP) were selected based on established literature demonstrating the feasibility of extracting these parameters from extraoral photographs and 3D scans [9–12]. The output variables corresponded to traditional cephalometric parameters, including angular measurements such as Sella-Nasion-A point angle (SNA), Sella-Nasion-B point angle (SNB), A point-Nasion-B point angle (ANB), and linear measurements such as the distances from Point A to Nasion (NP2PA), Pogonion to Nasion (NP2PO), Condylion point A distance (COPAD), and Condylion point B distance (COPOD). Input and output data are presented in Table 1.

The training dataset, comprising 1255 subjects from a diverse multiethnic population across the USA, underwent initial cleaning to analyse any potential missing data

Table 1 Input variables and output variables for cephalometric analysis

Input Values	Output Values
OB (Overbite): vertical overlap of the maxillary incisors over mandibular incisors.	SNA (Sella-Nasion-A point angle): angle determining anteroposterior position of maxilla.
OJ (Overjet): horizontal overlap of the maxillary incisors beyond the mandibular incisors.	SNB (Sella-Nasion-B point angle): angle determining anteroposterior position of mandible.
U6L6D (Upper 6 to Lower 6 Distance): horizontal measurement between upper and lower first molars	NP2PA (Point A to Nasion): linear measure indicating anteroposterior maxillary positioning.
LFHNP (LowerFaceHeight): vertical distance from Nasion to Pogonion.	NP2PO (Pogonion to Nasion): linear measure for mandibular position.
IIANG (InterIncisalAngle): angle formed between the long axes of upper and lower incisors.	SNDST (Sella-Nasion Distance): linear measure for the anterior cranial base.
TFHNP (TotalFaceHeight): overall vertical facial measurement from Nasion to Pogonion.	SNFHA (SN to Frankfort Angle): angle formed by the intersection of the Sella-Nasion line and the Frankfort plane.
PLHNP (Percent Lower Face Height): proportion of the lower face height relative to total face height.	CONVX (Angle of Convexity)
	AB2FH (AB Distance): linear distance between the projections of points A and B onto the Frankfort plane.
	U1SNA (Upper 1 to Sella-Nasion Angle)
	FMPA (Frankfort-Mandibular Plane): angle between Frankfort and mandibular planes.
	SADLA (Saddle Angle): angle formed by Nasion, Sella, and Articulare point.
	WITSA (Wits Appraisal): sagittal skeletal discrepancy based on dental occlusal plane.
	IMPA (L1 to Mandibular Plane Angle): inclination of L1 regarding the mandibular plane.
	ANB (A point-Nasion-B point angle): difference between SNA and SNB angles, indicating skeletal discrepancy.
	COPAD (Condylion point A distance): linear distance from Condylion to point A.
	COPOD (Condylion point B distance): linear distance from Condylion to point B.

points. Categorical data were encoded using one-hot encoding.

An initial exploratory data analysis (EDA) was conducted to ensure the robustness of the dataset, involving calculation and analysis of descriptive statistics (mean, standard deviation, range) and a correlation matrix to understand interrelationships among variables. The starting correlation matrix revealed significant correlations between key parameters, particularly among sagittal (SNA, SNB, and ANB) and vertical (TFHNP, LFHNP) measurements, underpinning the rationale for feature inclusion and model complexity.

Several regression algorithms were explored to develop predictive models, each rigorously optimised through GridSearchCV to ensure maximal performance as measured by the mean squared error (MSE). Models tested included ExtraTreesRegressor, CatBoostRegressor, SVR, GaussianProcessRegressor, Ridge, Lasso, and Bayesian Ridge, among others. The final selection of algorithms was based on performance metrics indicating minimal prediction errors [13–15].

A repeated random sub-sampling validation methodology was applied, repeatedly partitioning the dataset into training (80%, 1004 cases) and validation (20%, 251 cases) subsets. Additionally, an independent external validation dataset, consisting of 51 untreated orthodontic cases from the University of Insubria Dental Clinic, was employed to test the clinical applicability of the predictive models. Patients for this validation phase were randomly selected using the Random.org tool. Lateral clinical photographs were captured and analysed with WebCeph software, while 3D intraoral scans were obtained using the iTero scanner and analysed in STL format with MeshMixer software.

Cephalometric landmarks generated by AI predictions were meticulously reviewed and manually adjusted by experienced orthodontists to ensure accuracy and clinical relevance [16].

Performance evaluation involved comprehensive statistical analyses, including calculation of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Spearman correlation coefficient, R-squared statistics, and paired t-tests. These metrics provided insights into the accuracy, consistency, and statistical validity of the predictive models.

Results

The study's descriptive statistics showed a balanced gender distribution with an average patient age of 11.2 ± 3.8 years. A correlation analysis revealed significant relationships among numerous cephalometric variables, notably between SNA, SNB, and ANB, indicative of their

anatomical dependencies, and among vertical facial measurements, including TFHNP and LFHNP.

A comparative analysis using Mean Squared Error (MSE) showed significant variability among regression models, depending on the cephalometric parameter. Overall, the ExtraTreesRegressor consistently outperformed other models. CatBoostRegressor and SVR also delivered strong results for multiple angular and linear measurements. Conversely, GaussianProcessRegressor showed the highest error values across nearly all variables. Linear models, such as Ridge, Lasso, and BayesianRidge, displayed more variability and were generally outperformed by tree-based ensemble methods and neural networks (Table 2).

Error metrics for validation highlighted varying accuracy among predicted parameters, with RMSE ranging notably from 2.317 mm for NP2PA to 12.121 mm for COPAD (Table 4). Corresponding MAE values similarly ranged from a low of 1.886 mm (NP2PA) to a high of 10.867 mm (COPAD), emphasising differences in predictive reliability across measurements.

Despite relatively low average prediction errors for parameters like ANB (RMSE: 2.772°), the predictive models displayed limited variance explanation, as evidenced by R-squared values nearing zero. This pattern underscored the models' tendencies towards mean-centric predictions rather than capturing individual variations.

Spearman correlation analysis further nuanced the assessment, revealing high rank-order consistency in predictions of COPOD (0.850), COPAD (0.568), and U1SNA (0.548). Conversely, parameters like NP2PO exhibited weak negative correlations (–0.052), indicating poor predictive consistency. Table 3 reports the correlation coefficient derived from Spearman tests.

In the subsequent phase, we utilised the best-performing models from the initial testing to make predictions on a new set of 51 cases not included in the training dataset (Table 4).

Despite both aiming to assess the agreement between predictions and actual data, R-squared and Spearman's rank correlation capture different aspects of model performance. R-squared measures how much of the total variance is explained, while Spearman's correlation assesses whether the relative rank order is preserved. Moreover, various machine learning models, which may not rely on linear assumptions, can exacerbate these discrepancies. As a result, a model might register a low R-squared (poorly explaining variance) yet maintain a moderate Spearman correlation if it still ranks predictions in a manner that aligns reasonably well with the actual data (Table 4).

The paired t-test indicated statistically significant differences ($p < 0.05$) between predicted and actual measurements for some parameters, reflecting systematic biases in predictions. While this underscores the need for further model refinement, parameters with smaller and non-significant differences support potential clinical reliability.

Collectively, the CEPHCLINIC software demonstrated substantial predictive capabilities with varied accuracy dependent on specific cephalometric parameters. The results emphasise the need for further model refinement and dataset expansion, particularly for improved handling of parameters with weaker predictive accuracy.

Figure 1 presents the Bland-Altman plots for all predicted variables in comparison to their corresponding real values. For each parameter, the plot shows the mean difference (bias) and the limits of agreement, calculated as the mean difference ± 1.96 times the standard deviation of differences.

Discussion

The integration of AI into orthodontic diagnostics, specifically for predicting cephalometric measurements from clinical photographs and three-dimensional dental scans, represents a substantial advancement toward more efficient and safer diagnostic practices [17, 18]. Traditional cephalometric analyses, dependent on radiographic images, pose inherent health risks due to radiation exposure, especially significant in paediatric populations. The CEPHCLINIC software addresses these risks by significantly reducing reliance on radiographs, thus aligning with radioprotection principles such as ALADAIP, which advocate minimising radiation exposure without compromising diagnostic accuracy.

The present findings revealed considerable variation in predictive accuracy among different cephalometric parameters. The ExtraTreesRegressor algorithm consistently outperformed other models across multiple parameters, underscoring the strength of ensemble tree-based methods in managing complex, nonlinear relationships within clinical datasets. CatBoostRegressor and Support Vector Regression also showed commendable predictive capabilities, whereas GaussianProcessRegressor yielded notably inferior results, highlighting that algorithm selection critically influences predictive performance.

Statistical evaluation identified parameters such as ANB, COPOD, COPAD, and U1SNA as being more consistently predictable. For instance, COPOD displayed a robust Spearman correlation (0.850), emphasising strong rank-order prediction consistency despite its relatively higher absolute prediction errors. Conversely, parameters such as NP2PO and NP2PA presented significant challenges, indicated by weaker correlations and higher mean errors, necessitating further optimisation. These discrepancies likely reflect the inherent complexities in capturing vertical and angular relationships from surface-based morphological data compared to direct radiographic

Table 2 Mean squared error (MSE) of predictive models for each cephalometric value

Feature	Models	HistGradientBoostingRegressor	RandomForestRegressor	ExtraTrees-Regressor	Decision-TreeR-egres-sor	SVR	LinearS-VR	Ridge	Lasso	Elas-tic-Net	Lars	Or-thogo-nal-Match-ing-Pursuit	KNeigh-borsRe-gressor	XG-BRe-gres-sor	Cat-Boost-Regres-sor	ML-PRe-gres-sor	Bag-gin-gRe-gres-sor	Ada-Boost-Regres-sor	Pas-sive-Ag-gres-siveR-egres-sor	RANSACRe-gressor	Bayes-ian-Ridge	Gauss-ianPro-cessRe-gressor	Gra-dient-Boost-ingRe-gres-sor
SNA (°)		9.88	8.94	8.10	13.44	9.06	10.66	10.71	10.58	10.60	10.73	10.54	8.35	9.71	8.58	11.90	8.91	10.16	14.09	10.43	10.59	345.21	9.67
SNB (°)		6.55	5.36	4.78	9.38	5.40	6.47	6.46	6.32	6.33	6.48	6.69	4.83	6.05	5.22	6.00	5.23	6.36	8.30	6.32	6.43	302.51	6.13
ANB (°)		3.49	3.39	3.22	5.15	3.77	4.09	4.09	4.06	4.04	4.11	4.09	3.07	3.60	3.26	3.71	3.44	3.93	5.05	4.18	4.05	4.88	3.66
NP2PA (mm)		10.78	10.02	9.53	16.17	9.92	11.50	11.58	11.86	11.85	11.61	12.16	10.94	10.76	10.19	11.28	10.06	11.66	10.60	11.20	11.38	11.47	10.55
NP2PO (mm)		27.10	28.50	26.74	48.75	26.36	32.78	32.75	33.68	34.35	32.86	35.03	30.08	29.01	27.18	29.09	28.92	34.00	33.71	33.06	32.77	34.80	29.20
SNDST (mm)		9.12	8.50	8.72	11.70	8.83	12.21	12.19	12.57	12.69	12.19	12.30	7.92	9.92	8.55	9.45	8.56	11.37	12.87	12.31	12.30	266.77	9.82
SNFHA (°)		10.04	10.00	9.67	15.30	9.42	12.15	12.20	12.40	12.18	13.04	12.42	10.55	10.66	9.34	11.02	10.08	11.36	13.55	12.94	12.03	12.71	10.52
CONVX (°)		20.75	18.86	17.68	30.84	20.08	23.24	23.11	22.54	22.51	23.15	22.66	17.89	21.50	17.92	22.08	18.68	21.76	23.26	23.79	22.63	1392.72	20.41
AB2FH (mm)		8.37	8.80	7.76	13.51	7.84	10.12	10.13	10.41	10.45	10.12	10.62	8.26	8.68	7.75	8.34	8.91	10.16	10.51	10.28	10.16	12.14	9.08
U1SNA (°)		15.56	14.24	13.60	24.45	12.66	15.90	15.84	16.02	15.97	15.90	17.11	13.71	15.90	14.00	13.40	14.62	16.83	17.35	16.26	15.82	312.73	15.66
FMPA (°)		17.36	16.01	14.77	27.11	15.71	19.26	19.27	19.45	19.29	19.28	19.80	16.25	16.57	14.92	14.86	16.34	18.81	23.26	19.37	19.23	49.92	15.95
SADLA (°)		16.37	14.42	14.31	24.76	15.45	20.10	20.03	20.34	20.29	20.01	20.21	15.58	16.74	14.59	18.04	14.84	18.63	21.66	20.08	20.28	880.14	17.00
WITSA (mm)		0.85	0.85	0.78	1.38	0.76	0.74	0.74	0.74	0.74	0.74	0.73	0.94	0.77	0.71	0.72	0.84	0.88	1.00	0.76	0.74	1.43	0.79
IMPA (°)		14.34	13.72	12.68	21.17	15.29	16.34	16.20	16.97	17.08	16.14	17.01	15.30	14.64	13.74	13.92	13.55	18.15	20.93	15.87	16.30	472.78	14.71
COPAD (mm)		16.90	15.70	15.14	22.71	15.46	17.61	17.67	17.83	18.02	17.62	17.64	15.36	15.43	14.34	13.92	15.64	17.08	21.55	19.31	17.78	415.58	15.58
COPOD (mm)		14.14	14.43	13.14	22.49	15.03	14.86	14.93	14.85	15.00	14.96	14.92	14.35	14.32	13.55	13.49	14.48	16.92	17.93	15.20	14.93	613.18	14.68

measurements. Therefore, the current findings underline that specific cephalometric parameters, notably COPOD, COPAD, ANB, and U1SNA, exhibit greater predictive reliability and thus hold immediate potential for clinical implementation. Conversely, predictions for NP2PO and similar measurements necessitate further refinement and should currently be used with caution in clinical settings.

A critical observation in our study was the discrepancy between low RMSE and minimal variance explanation (low R-squared values), particularly evident with ANB. This indicates that while the predictions were generally accurate in average terms, they failed to capture patient-specific variability adequately. Consequently, while CEPHCLINIC shows promise for clinical use, it requires further refinement to enhance precision in individual diagnostic contexts.

Although the prediction of ANB (RMSE: 2.772°) may seem small in absolute terms, it is important to recognise that even differences of 2° in this measure can affect skeletal classification. Nonetheless, this level of error is comparable to inter-operator variability and typical Dahlberg errors reported in cephalometric studies. Furthermore, in clinical practice, angular relationships like ANB are never interpreted on their own but are combined with complementary values such as SNA and SNB. Therefore, while these predictive inaccuracies should be acknowledged, they do not diminish the clinical usefulness of the model when used alongside professional judgment.

Overall, the results indicate that linear skeletal parameters, such as COPOD, COPAD, and SNDST, tended to produce higher performance metrics in terms of correlation and RMSE. However, they sometimes showed significant bias and broad limits of agreement. These parameters benefit from the inherent stability of skeletal landmarks and their typically well-defined geometric relationships on lateral cephalograms, making them more predictable from other anatomical structures.

In contrast, angular skeletal parameters, such as SNA, SNB, ANB, and FMPA, showed more heterogeneous results. While models for SNA and SNB showed acceptable bias levels and moderate agreement, the performance for ANB was marked by a near-zero bias but wide limits of agreement and a low R^2. This underscores how even minor errors in predicting angular measures can cause significant variability, likely due to the combined effect of angle sensitivity to landmark perturbations. FMPA and CONVX also illustrate this issue, showing relatively good bias but considerable dispersion in predictions. These findings suggest that although angular skeletal measurements may be crucial for diagnostic frameworks, their predictability is influenced by the indirect nature of angle formation from multiple reference points.

The predictive performance of the models shows considerable variation across different types of cephalometric parameters. Notably, the distinction between skeletal and dental, as well as linear and angular measures, does not follow a straightforward pattern of predictability. Among linear skeletal parameters, COPOD and COPAD demonstrated the strongest performance, with Spearman correlations of 0.850 and 0.568, respectively. These results suggest that external mandibular morphology, which is easily visible in soft tissue profiles, is effectively captured by the models. In contrast, other linear skeletal parameters such as NP2PA and NP2PO showed negligible or even negative correlations, despite their low RMSE values. This discrepancy indicates that the models often regress towards the mean when external landmarks are absent, especially for measures involving deep cranial base structures (e.g., the Sella-Nasion line). The performance of angular skeletal parameters was generally modest. Measures like SNA, SNB, and ANB yielded very low correlations (ranging from −0.182 to 0.205), implying that the models find it difficult to reconstruct cranial angular relationships from visual data alone. Conversely, dentoskeletal angular parameters such as U1SNA and IMPA achieved some of the highest correlations (0.548 and 0.391, respectively). This better performance may be due to the visibility of perioral soft tissues, which seem to provide sufficient cues for estimating incisor inclination and orientation, even without direct radiographic information.

Overall, the analysis shows that predictability relies less on whether a parameter is linear or angular and more on whether the relevant anatomical structures are externally visible. Parameters associated with landmarks that can be inferred from the facial surface, especially the mandible and the anterior dental region, tend to demonstrate better predictive agreement. Conversely, those involving deep skeletal landmarks, particularly in the cranial base, remain difficult for surface-based inference models.

. Although the average bias for these was generally small or non-significant, the limits of agreement were very wide, indicating a lack of reliability in individual predictions. These differences might be due to the higher biological variability of dental positions, as well as the influence of treatment history, occlusal function, or dental compensations, which add more unpredictability. Furthermore, minor errors in landmarking dental points, particularly when the incisor edges or apices are poorly defined, can significantly affect angular calculations, increasing the inaccuracy.

Interestingly, even angular parameters derived from skeletal landmarks, such as NP2PO and SADLA, showed poor performance, with significant biases and very wide agreement intervals. This indicates that angular variables, regardless of their anatomical origin, are generally

Table 3 Correlation matrix based on spearman correlation test

	Sex	Age	WITSA	OB_FH	OJ_OC	U6L6D	LFHNP	IIANG	IMPA	COPAD	COPOD	SNA	SNB	ANB	NP2PA	NP2PO	SNDST	SNFHA	CONVX	AB2FH	TFHNP	PLHNP	U1SNA	FMPA	SADLA
Sex	-	−0.02	−0.03	0.16	0.14	−0.05	−0.12	−0.06	−0.05	−0.18	−0.19	−0.11	−0.19	0.10	−0.02	0.02	−0.29	0.11	−0.04	0.06	−0.15	−0.08	0.02	0.09	0.27
Age	−0.02	-	0.19	−0.18	−0.16	0.22	0.32	−0.07	0.23	0.50	0.70	0.02	0.24	−0.29	0.25	0.00	0.37	−0.26	0.38	−0.06	0.58	−0.12	−0.13	−0.08	−0.09
WITSA	−0.03	0.19	-	−0.34	−0.35	0.89	0.09	−0.08	−0.08	0.07	0.25	−0.01	0.29	−0.38	0.17	−0.06	0.07	−0.16	0.32	−0.27	0.15	0.03	−0.19	0.11	−0.12
OB_FH	0.16	−0.18	−0.34	-	0.97	−0.37	0.03	−0.15	−0.09	0.15	−0.07	0.01	−0.40	0.53	−0.22	0.13	0.10	0.22	−0.36	0.44	0.02	0.05	−0.11	0.07	0.18
OJ_OC	0.14	−0.16	−0.35	0.97	-	−0.38	0.01	−0.09	−0.09	0.18	−0.07	0.01	−0.40	0.53	−0.15	0.20	0.14	0.15	−0.36	0.50	0.02	−0.01	−0.05	0.10	0.16
U6L6D	−0.05	0.22	0.89	−0.37	−0.38	-	0.11	−0.08	−0.06	0.12	0.30	0.01	0.31	−0.39	0.15	−0.07	0.11	−0.15	0.32	−0.28	0.19	0.03	−0.18	0.10	−0.09
LFHNP	−0.12	0.32	0.09	0.03	0.01	0.11	-	−0.07	0.32	0.38	0.46	−0.04	0.04	−0.10	0.04	−0.02	0.39	0.01	0.15	−0.01	0.88	0.69	−0.07	−0.28	−0.03
IIANG	−0.06	−0.07	−0.08	−0.15	−0.09	−0.08	−0.07	-	−0.60	−0.09	−0.07	−0.04	0.01	−0.08	−0.01	−0.16	−0.04	0.05	0.07	−0.13	−0.09	−0.10	0.74	−0.25	−0.05
IMPA	−0.05	0.23	−0.08	−0.09	−0.09	−0.06	0.32	−0.60	-	0.18	0.10	0.18	0.10	0.13	0.00	0.11	0.12	−0.17	−0.07	0.16	0.26	0.19	−0.29	−0.29	0.00
COPAD	−0.18	0.50	0.07	0.15	0.18	0.12	0.38	−0.09	0.18	-	0.88	0.23	0.18	0.08	−0.02	0.05	0.82	−0.15	0.03	0.21	0.67	−0.07	−0.09	−0.03	0.02
COPOD	−0.19	0.70	0.25	−0.07	−0.07	0.30	0.46	−0.07	0.10	0.88	-	0.12	0.32	−0.26	0.16	−0.02	0.77	−0.24	0.34	−0.04	0.78	0.00	−0.18	−0.01	−0.08
SNA	−0.11	0.02	−0.01	0.01	0.01	0.01	−0.04	−0.04	0.18	0.23	0.12	-	0.71	0.47	−0.49	−0.15	−0.08	−0.42	−0.46	0.20	−0.06	0.08	−0.18	−0.02	−0.41
SNB	−0.19	0.24	0.29	−0.40	−0.40	0.31	0.04	0.01	0.10	0.18	0.32	0.71	-	−0.30	−0.13	−0.36	0.00	−0.53	0.18	−0.40	0.05	0.06	−0.41	−0.22	−0.50
ANB	0.10	−0.29	−0.38	0.53	0.53	−0.39	−0.10	−0.08	0.13	0.08	−0.26	0.47	−0.30	-	−0.50	0.23	−0.11	0.09	−0.86	0.76	−0.15	0.04	0.27	0.24	0.06
NP2PA	−0.02	0.25	0.17	−0.22	−0.15	0.15	0.04	−0.01	0.00	−0.02	0.16	−0.49	−0.13	−0.50	-	0.69	0.21	−0.58	0.49	0.15	0.16	−0.20	−0.06	0.33	−0.09
NP2PO	0.02	0.00	−0.06	0.13	0.20	−0.07	−0.02	−0.16	0.11	0.05	−0.02	−0.15	−0.36	0.23	0.69	-	0.17	−0.56	−0.21	0.78	0.09	−0.17	0.12	0.66	−0.06
SNDST	−0.29	0.37	0.07	0.10	0.14	0.11	0.39	−0.04	0.12	0.82	0.77	−0.08	0.00	−0.11	0.21	0.17	-	−0.11	0.18	0.15	0.65	−0.03	−0.05	−0.02	−0.08
SNFHA	0.11	−0.26	−0.16	0.22	0.15	−0.15	0.01	0.05	−0.17	−0.15	−0.24	−0.42	−0.53	0.09	−0.58	−0.56	−0.11	-	−0.07	−0.33	−0.08	0.13	0.24	−0.32	0.48
CONVX	−0.04	0.38	0.32	−0.36	−0.36	0.32	0.15	0.07	−0.07	0.03	0.34	−0.46	0.18	−0.86	0.49	−0.21	0.18	−0.07	-	−0.61	0.22	0.00	−0.24	−0.26	−0.02
AB2FH	0.06	−0.06	−0.27	0.44	0.50	−0.28	−0.01	−0.13	0.16	0.21	−0.04	0.20	−0.40	0.76	0.15	0.78	0.15	−0.33	−0.61	-	0.07	−0.08	0.25	0.54	−0.03
TFHNP	−0.15	0.58	0.15	0.02	0.02	0.19	0.88	−0.09	0.26	0.67	0.78	−0.06	0.05	−0.15	0.16	0.09	0.65	−0.08	0.22	0.07	-	0.35	−0.04	−0.11	−0.03
PLHNP	−0.08	−0.12	0.03	0.05	−0.01	0.03	0.69	−0.10	0.19	−0.07	0.00	0.08	0.06	0.04	−0.20	−0.17	−0.03	0.13	0.00	−0.08	0.35	-	−0.13	−0.23	−0.08
U1SNA	0.02	−0.13	−0.19	−0.11	−0.05	−0.18	−0.07	0.74	−0.29	−0.09	−0.18	−0.18	−0.41	0.27	−0.06	0.12	−0.05	0.24	−0.24	0.25	−0.04	−0.13	-	0.01	0.18
FMPA	0.09	−0.08	0.11	0.07	0.10	0.10	−0.28	−0.25	−0.29	−0.03	−0.01	−0.02	−0.22	0.24	0.33	0.66	−0.02	−0.32	−0.26	0.54	−0.11	−0.23	0.01	-	−0.07
SADLA	0.27	−0.09	−0.12	0.18	0.16	−0.09	−0.03	−0.05	0.00	0.02	−0.08	−0.41	−0.50	0.06	−0.09	−0.06	−0.08	0.48	−0.02	−0.03	−0.03	−0.08	0.18	−0.07	-

Table 4 Statistical analysis outcomes: comparative metrics of true versus predicted data values across various features (test set, 51 cases)

Feature	RMSE	MAE	*R*-squared	Spearman Correlation	Paired t-test *p*-value
SNA	5.363	4.538	0.008	−0.182	0.000
SNB	3.965	3.298	0.015	0.118	0.008
ANB	2.772	2.241	0.000	0.205	0.010
NP2PA	2.317	1.886	0.000	−0.032	0.794
NP2PO	9.049	8.814	0.011	−0.052	0.000
SNDST	10.292	9.605	0.220	0.385	0.000
SNFHA	3.152	2.707	0.026	0.123	0.206
CONVX	7.200	5.630	0.000	0.089	0.014
WITSA	3.338	2.695	0.212	−0.429	0.330
AB2FH	4.982	4.079	0.010	−0.037	0.002
IMPA	6.305	4.961	0.199	0.391	0.230
U1SNA	6.670	5.256	0.328	0.548	0.099
COPAD	12.121	10.867	0.304	0.568	0.000
COPOD	11.532	10.620	0.645	0.850	0.000
FMPA	5.303	4.396	0.097	0.257	0.066
SADLA	9.223	8.028	0.082	0.464	0.000

less reliable due to their mathematical reliance on accurate landmark triangulation. Conversely, linear dental parameters were underrepresented but displayed variable behaviour; for example, COPAD (which relates to molar anteroposterior position) demonstrated better correlation than expected, although it was still influenced by systematic error.

Taken together, these findings suggest a systematic pattern: linear parameters, especially those involving skeletal landmarks, are generally more predictable and stable, whereas angular and dental parameters tend to be more affected by variability, likely due to a mix of biological complexity and geometric amplification of landmarking uncertainty.

Beyond classical performance metrics, the clinical applicability of the models was further evaluated through Bland-Altman plots, which assess agreement between real and predicted values.

The Bland-Altman analysis provides a comprehensive view of the agreement between real and predicted values for the evaluated cephalometric parameters, highlighting both the strengths and limitations of the prediction models. An apparent heterogeneity appears among the different features. Notably, NP2PA stands out as the parameter with the best overall agreement, showing an almost zero bias and narrow limits of agreement. This indicates that the corresponding model attains a high level of accuracy and precision, supporting its potential reliability in clinical settings. Similarly, SNFHA, U1SNA, and IMPA show minimal or non-significant bias. However, the range of the agreement intervals remains wide in some cases, implying that while the central tendency of predictions may be appropriate, the variability in individual estimates could limit clinical utility.

In contrast, other parameters such as COPAD, COPOD, and SNDST reveal notable systematic errors. These models consistently underestimate or overestimate actual values, and the differences are not only statistically significant but also quite large. The width of the agreement intervals in these cases is considerable, often exceeding 15 units, which indicates low precision and makes the model unsuitable for use in precise diagnostic or treatment planning workflows. Interestingly, despite COPOD exhibiting the highest Spearman correlation among all parameters (0.850), the associated Bland-Altman plot demonstrates that such correlation does not imply clinical agreement. This observation emphasises the importance of complementing correlation-based metrics with agreement-based ones to achieve a more realistic assessment of model performance.

Parameters such as ANB, FMPA, and WITSA occupy an intermediate position. They exhibit low or negligible bias but still show wide limits of agreement. These models may thus provide acceptable average predictions across populations but are less reliable at the individual level. Particularly in ANB, the minimal bias accounts for the relatively low RMSE, yet the poor R^2 indicates that the model fails to capture the data's variance, a finding reflected in the considerable spread seen in the Bland-Altman plot.

A subset of parameters, such as NP2PO and SADLA, exhibits poor agreement because of considerable bias and very wide limits. These models tend to consistently deviate from the true values, either by underestimating or overestimating, and this error is worsened by high variability, showing instability across cases. Their clinical use would be risky as they do not offer reliably consistent estimations.

Models like AB2FH and CONVX exhibit statistically significant biases and wide confidence intervals, indicating that although their predictions may be systematically biased, they are also highly variable. These characteristics highlight not only model miscalibration but also a lack of robustness across the data distribution.

The present research also highlighted significant demographic considerations. The training dataset predominantly comprised a North American population, contrasting with the primarily Italian validation group. This demographic difference introduces a critical limitation, as craniofacial characteristics vary notably across ethnicities, which may limit the generalizability of the CEPHCLINIC software to other ethnic groups. This variability underscores the importance of expanding future training datasets to include broader demographic representations, thereby enhancing the software's generalizability and clinical applicability globally.

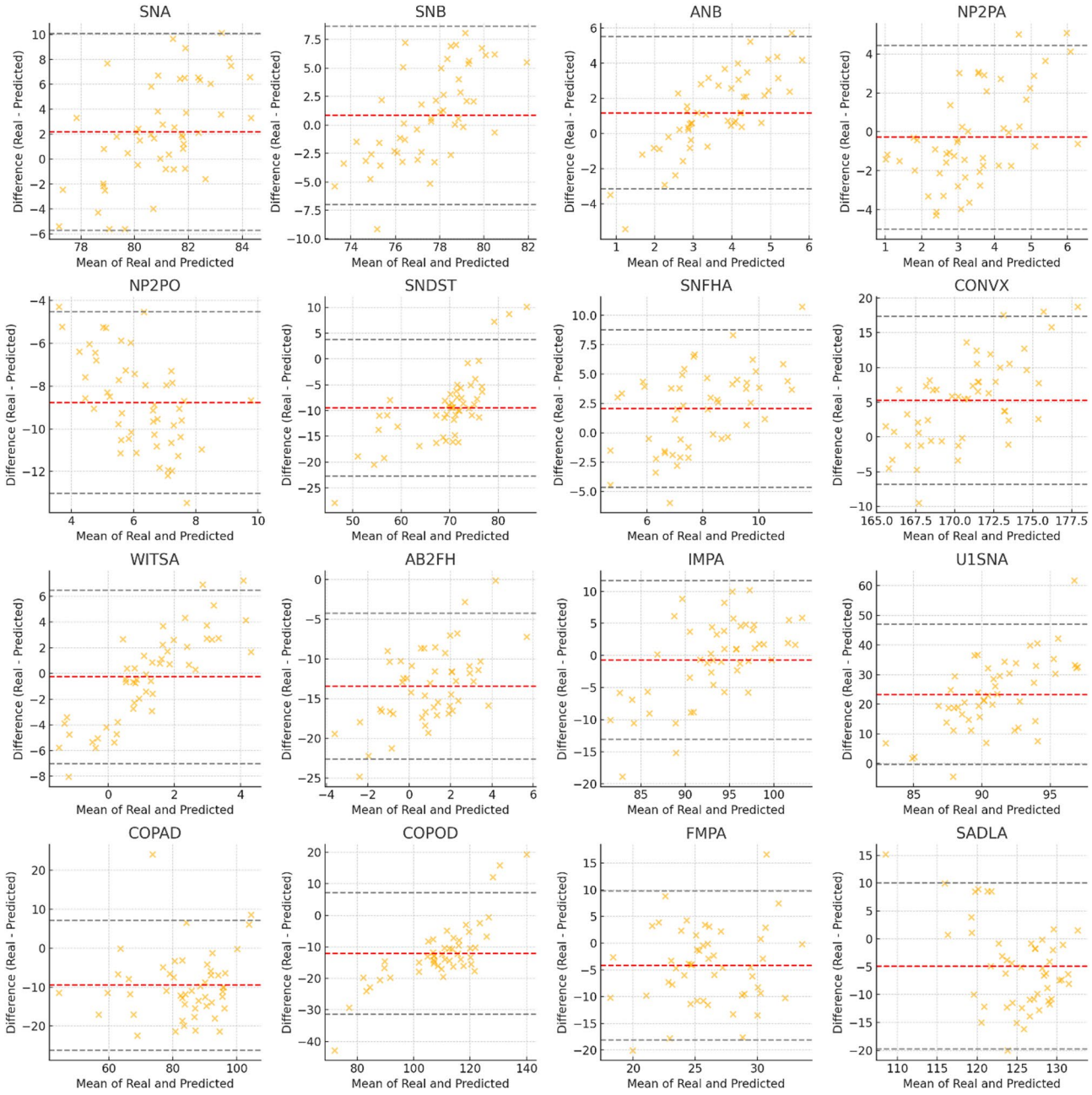

Fig. 1 Statistical analysis outcomes: comparative metrics of true versus predicted data values

While CEPHCLINIC significantly reduces exposure to ionising radiation, it does not substitute radiographic imaging entirely, especially in scenarios requiring detailed diagnostic evaluation, such as detecting root resorption, bone lesions, or other structural anomalies visible exclusively through traditional radiography. Thus, it should be perceived as complementary to radiographic methods rather than a complete replacement.

Conclusions

The CEPHCLINIC software represents a promising adjunct to traditional cephalometric analyses, effectively reducing radiation exposure without significantly compromising diagnostic reliability. Its utility, however, hinges upon further optimisation, broader demographic representation, and integration with complementary diagnostic technologies. Continued research and refinement will likely augment its effectiveness, ultimately contributing to safer and more personalised orthodontic diagnostics.

Future enhancements of the CEPHCLINIC software should focus on refining predictive algorithms, incorporating more extensive and diverse datasets, and optimising the user interface to facilitate routine clinical integration. Additionally, integration with existing digital orthodontic tools, such as intraoral scanners and 3D analysis software, could significantly enhance its clinical utility.

Acknowledgements
The authors would like to thank the staff and volunteers of the Dental Clinic at the University of Insubria for their assistance in collecting the data. We are also grateful to the American Association of Orthodontists Foundation (AAOF) for providing access to their craniofacial growth dataset.

Authors' contributions
PAZ- ConceptualizationMC- ValidationLL- Data CurationAC- Writing - Original DraftMS- Writing - Review & EditingPRB- SupervisionMR- Project administration.

Funding
This work did not receive any specific grant from public, commercial, or not-for-profit funding agencies. The authors confirm that no external or third-party funding was used to support this research.

Data availability
The datasets supporting the conclusions of this article are available from the corresponding author upon reasonable request. No public repository currently hosts the clinical data used in this study due to privacy considerations.

Declarations

Ethics approval and consent to participate
This study was approved by the University of Insubria Ethics Committee (protocol number 0026262 of 21/02/2025). The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. All participants provided informed consent before enrolment in the study.

Consent for publication
No individual person's data (including personal details, images, or videos) is presented in this manuscript. Therefore, consent for publication is not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]DIMIT, Department of Medicine and Technological Innovation, University of Insubria, Varese, Italy
[2]School of Medicine, University of Insubria, Varese, Italy
[3]DISUIT, Department of Human Sciences, Innovation and Territory, University of Insubria, Varese, Italy
[4]DBSV, Department of Biotechnology and Life Sciences, University of Insubria, Varese, Italy
[5]Translational Medicine, Department of Biomedical Sciences for Health, University of Milan, Milan, Italy

Received: 8 April 2025 / Accepted: 4 September 2025
Published online: 03 October 2025

References

1. Polizzi A, Boato M, Serra S, D'Antò V, Leonardi R. Applications of artificial intelligence in orthodontics: a bibliometric and visual analysis. Clin Oral Investig. 2025;29:65. https://doi.org/10.1007/S00784-025-06158-Y/FIGURES/7.
2. Dinesh A, Mutalik S, Feldman J, Tadinada A. Value-addition of lateral cephalometric radiographs in orthodontic diagnosis and treatment planning. Angle Orthod. 2020;90:665–71. https://doi.org/10.2319/062319-425.1.
3. Linet MS, Kim KP, Rajaraman P. Children's exposure to diagnostic medical radiation and cancer risk: epidemiologic and dosimetric considerations. Pediatr Radiol. 2009;39:S4–26. https://doi.org/10.1007/S00247-008-1026-3.
4. Cho HJ, Lee S-S, Kang JH, Kim J-E, Huh K-H, Yi W-J, et al. Development of 10 principles of radiation protection in oral and maxillofacial radiology. Imaging Sci Dent. 2025. https://doi.org/10.5624/ISD.20250041.
5. Neeraja R, Anbarasi LJ. A critical review of artificial intelligence based techniques for automatic prediction of cephalometric landmarks. Artif Intell Rev. 2025;58:1–56. https://doi.org/10.1007/S10462-025-11135-8/TABLES/7.
6. Ingle NA, Alabsi NF, Al-Hashimi H, Albuolayan NA, Alburidy F, Alanazi F, Alhammad AT. The use of artificial intelligence in orthodontic treatment planning: A systematic review and Meta-analysis. Adv Hum Biology. 2025;15:158–66. https://doi.org/10.4103/AIHB.AIHB_140_24.
7. Zecca PA, Caccia M, Levrini L, Carganico A, Reguzzoni M, Donadio D, et al. Ai-based open-source software for cephalometric analysis from limited FOV radiographs. J Dent. 2024;151:105412. https://doi.org/10.1016/J.JDENT.2024.105412.
8. Cho SJ, Moon JH, Ko DY, Lee JM, Park JA, Donatelli RE, Lee SJ. Orthodontic treatment outcome predictive performance differences between artificial intelligence and conventional methods. Angle Orthod. 2024;94:557–65. https://doi.org/10.2319/111823-767.1.
9. Gupta S, Tandon P, Singh GP, Shastriv D. Comparative assessment of cephalometric with its analogous photographic variables. Natl J Maxillofac Surg. 2022;13:99–107. https://doi.org/10.4103/NJMS.NJMS_267_20.
10. Mehta P, Sagarkar RM, Mathew S. Photographic assessment of cephalometric measurements in skeletal class II cases: a comparative study. J Clin Diagn Res. 2017;11:ZC60–4. https://doi.org/10.7860/JCDR/2017/25042.10075.
11. Zhang X, Hans MG, Graham G, Kirchner HL, Redline S. Correlations between cephalometric and facial photographic measurements of craniofacial form. Am J Orthod Dentofac Orthop. 2007;131:67–71. https://doi.org/10.1016/j.ajodo.2005.02.033.
12. Zecca P P.A., Fastuca R, Beretta M, Caprioglio A, Macchi A. Correlation assessment between Three-Dimensional facial soft tissue scan and lateral cephalometric radiography in orthodontic diagnosis. Int J Dent. 2016;2016. https://doi.org/10.1155/2016/1473918.
13. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. accessed July 2, CatBoost: unbiased boosting with categorical features, (n.d.). https://github.com/catboost/catboost (2025).
14. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System, (n.d.). https://doi.org/10.1145/2939672.2939785
15. Van Der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-image: image processing in python. PeerJ. 2014. https://doi.org/10.7717/PEERJ.453.
16. Narkhede S, Rao P, Sawant V, Sachdev SS, Arora S, Pawar AM, et al. Digital versus manual tracing in cephalometric analysis: a systematic review and meta-analysis. J Pers Med. 2024. https://doi.org/10.3390/jpm14060566.
17. Mao B, Tian Y, Xiao Y, Li J, Zhou Y, Wang X. Classification of skeletal discrepancies by machine learning based on three-dimensional facial scans. Int J Oral Maxillofac Surg. 2025;54:747–56. https://doi.org/10.1016/J.IJOM.2025.03.003.
18. Vaughan M, Mheissen S, Cobourne M, Ahmed F. Diagnostic accuracy of artificial intelligence for dental and occlusal parameters using standardized clinical photographs. Am J Orthod Dentofacial Orthop. 2025. https://doi.org/10.1016/J.AJODO.2025.01.017.

Publisher's Note