Dentomaxillofacial Radiology (2022) 51, 20200548
© 2022 The Authors. Published by the British Institute of Radiology

birpublications.org/dmfr

SHORT COMMUNICATION

Artificial intelligence-based cephalometric landmark annotation and measurements according to Arnett's analysis: can we trust a bot to do that?

[1]Thaísa Pinheiro Silva, [2]Mariana Mendonça Hughes, [3]Liciane dos Santos Menezes,
[4]Maria de Fátima Batista de Melo, [5]Paulo Henrique Luiz de Freitas and [6]Wilton Mitsunari Takeshita

[1]Departament of Oral Diagnosis, Piracicaba Dental School, University of Campinas, Piracicaba, São Paulo, Brasil; [2]Department of Dentistry, Undergraduate student of Dentistry, Federal University of Sergipe, Sergipe, Brazil; [3]Department of Dentistry, Federal University of Bahia, Salvador, Brazil; [4]Department of Dentistry, PhD in Oral Radiology, Federal University of Sergipe, Sergipe, Brazil; [5]Department of Dentistry, PhD in Oral Life Sciences (OMF Surgery), Federal University of Sergipe at Lagarto, Sergipe, Brazil; [6]Department of Dentistry, PhD in Oral Radiology and Postdoctoral in Integrated Dentistry, Professor of Oral Radiology, Oral Diagnosis and Bioestatistics, Federal University of Sergipe, Sergipe, Brazil

Objective: To assess the reliability of CEFBOT, an artificial intelligence (AI)-based cephalometry software, for cephalometric landmark annotation and linear and angular measurements according to Arnett's analysis.
Methods: Thirty lateral cephalometric radiographs acquired with a Carestream CS 9000 3D unit (Carestream Health Inc., Rochester/NY) were used in this study. The 66 landmarks and the 10 selected linear and angular measurements of Arnett's analysis were identified on each radiograph by a trained human examiner (control) and by CEFBOT (RadioMemory Ltd., Belo Horizonte, Brazil). For both methods, landmark annotations and measurements were duplicated with an interval of 15 days between measurements and the intraclass correlation coefficient (ICC) was calculated to determine reliability. The numerical values obtained with the two methods were compared by a *t*-test for independent variables.
Results: CEFBOT was able to perform all but one of the 10 measurements. ICC values > 0.94 were found for the remaining eight measurements, while the Frankfurt horizontal plane - true horizontal line (THL) angular measurement showed the lowest reproducibility (human, ICC = 0.876; CEFBOT, ICC = 0.768). Measurements performed by the human examiner and by CEFBOT were not statistically different.
Conclusion: Within the limitations of our methodology, we concluded that the AI contained in the CEFBOT software can be considered a promising tool for enhancing the capacities of human radiologists.
Dentomaxillofacial Radiology (2022) **51**, 20200548. doi: 10.1259/dmfr.20200548

Cite this article as: Silva TP, Hughes MM, Menezes LS, de Melo MFB, Freitas PHL, Takeshita WM. Artificial intelligence-based cephalometric landmark annotation and measurements according to Arnett's analysis: can we trust a bot to do that?. *Dentomaxillofac Radiol* (2022) 10.1259/dmfr.20200548.

Keywords: Artificial Intelligence; Cephalometry; Machine learning; Deep learning; Orthognathic Surgery

Correspondence to: Dr Wilton Mitsunari Takeshita, E-mail: wmtakeshita2@gmail.com

Received 14 November 2020; revised 06 April 2021; accepted 10 April 2021; published online 05 August 2022

Introduction

Cephalometric landmark annotation and tracing is fundamental for sound orthodontic diagnosis. Originally, cephalometric landmark annotation was performed manually on acetate sheets. Since the advent of radiograph digitization, however, cephalometric tracing gradually switched to its digital, contemporary form. One of the reasons for such shift toward digital methods was the very time-consuming nature of manual cephalometric tracing.[1] To address the limitation of the manual methods, innovative software was created to identify skeletal and soft tissue landmarks and generate cephalometric measurements. With this digitallydriven increase in efficiency, cephalometry is even more an undisputed diagnostic tool for orthodontics, orthognathic surgery and craniofacial growth and development analysis.[2]

Indeed, it is safe to say that digital cephalometry is the current gold-standard. Digitally annotated cephalometric landmarks are not only superior to those obtained manually, but they are also easier to collect and record, demand less time from the professional and allow for the application of image-enhancing digital tools.[3] With the emergent applications of Artificial Intelligence (AI) in Dentistry, one may expect even faster tracing and measuring along with increased accuracy in landmark annotation.[4]

AI is the constellation of technologies (algorithms, robotics, neural networks) that allows a software to have human-like intelligence properties such as machine learning.[5] A natural application for AI-based software in Dentistry is the automatization of cephalometric landmarking and tracing, which may reduce the time and subjectivity that sometimes compromise proper determination of cephalometric points.[6]

CEFBOT is a machine learning (ML)-based software that estimates the position of predefined landmarks (anatomic points) in a digital dental radiograph. The software was trained to locate 96 different points in a cephalometric image, covering the most used cephalometric analysis worldwide. CEFBOT predictions are performed by three concomitant subsystems that work to determine the anatomical point coordinates and a fourth, "quality control" subsystem. The first is a set of pipelines of statistical, geometrical, and deep learning (CNN)-based processes that transform the original image in a sequence of probability maps of the points' locations. The second subsystem, a CNN-based process, segments the original image by extracting and vectorizing the borders between bones and soft tissues of interest and, after that, infers the geometric location of main skull landmarks. The third subsystem is a ML and geometric process pipeline which combines the two previous results, converting the computerized information into point coordinates. The fourth subsystem uses yet another ML algorithm to analyze the final set of coordinates and validate the combined geometric coherence. The ML algorithms have been trained with approximately 250.000 cephalometric exams performed and annotated by professionals, thus covering a variety of patient profiles and digital X-ray acquisition devices.[7]

While there are several cephalometric analyses available, Arnett's analysis stands out for its worldwide use and its status of standard for orthognathic surgical diagnosis and planning.[8] Despite its clinical and scientific prominence, Arnett's cephalometric analysis has not, to the best of our knowledge, been performed with the aid of an AI-based software.

Therefore, the aim of this study is to determine whether the automated Arnett's cephalometric landmark annotation and the accompanying linear and angular performed by CEFBOT can be considered reliable for clinical and research purposes.

Methods and materials

This study was registered and approved by the Human Research Ethics Committee of the Federal University of Sergipe (UFS) University Hospital (CAAE: 28222720.5.0000.5546, committee opinion #3.852.687) and was carried out according to the STROBE initiative and the Declaration of Helsinki.

Sampling

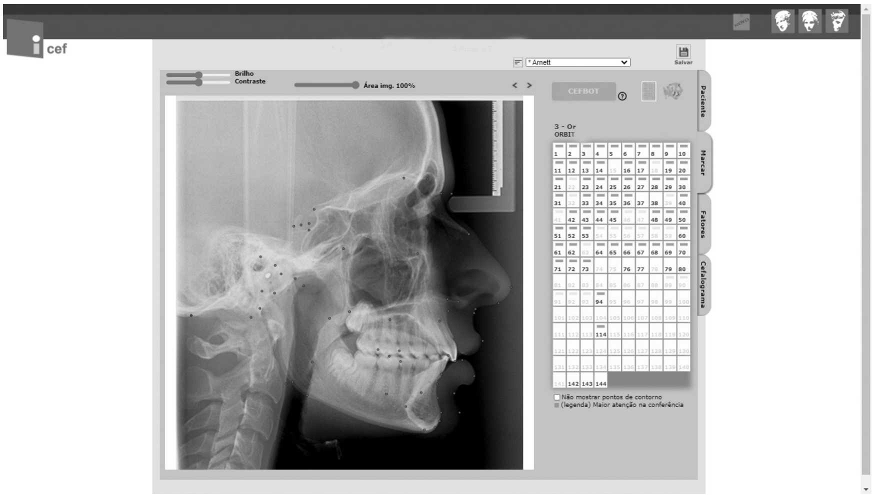
Sample size was calculated considering an intraclass correlation coefficient (ICC) of 0.70, 99% test power and 5% significance level, which resulted in the need for 28 lateral cephalometric radiographs.[9]

Lateral cephalometric radiographs were selected from the image bank of the Dental Radiology chair (UFS Dental School at Aracaju, Brazil). The selected images had been acquired with a Carestream CS 9000 3D unit (Carestream Health Inc, Rochester/NY) from individuals of both sexes, who were above 18 years of age and Angle class III (ANB, or A-point-nasion-B-point angle,≤0°). Radiographs with poor head positioning, as well as those from subjects with severe craniofacial deformities or facial asymmetries, with unerupted incisors and/or unerupted teeth covering the apexes of the incisors were excluded.

The 30 lateral cephalograms used in the study were saved as JPEG image files, with maximum dimensions of 1360 × 1840 pixels and resolution of 600 dpi, and then stored and examined on a personal computer MacBook Air 13-inch (Apple, Intel Core i5 processor, 8 GB memory, 256 GB storage, Intel HD Graphics 6000, and LED display).

Cephalometric analysis

Examiner calibration prior to cephalometric analyses was done using 30% of the sample. A dental radiologist



□ Não mostrar pontos de contorno
▨ (legenda) Maior atenção na conferência

Figure 1 Cephalometric landmarking according to Arnett's analysis performed by CEFBOT

with more than 20 years of experience in computerized cephalometry performed the landmarking repeatedly until an ICC >0.90 was achieved.

The calibrated examiner was responsible for identifying the 66 cephalometric points of Arnett's analysis,[8] which is commonly used by oral and maxillofacial surgeons when planning orthognathic surgery. Ten linear and angular measurements were calculated with the aid of RadioCef Studio 3 software, (RadioMemory Ltd., Belo Horizonte, Brazil). The numerical values obtained by the human examiner were regarded as control and recorded on a Numbers (Apple Inc., Cupertino, Califórnia) spreadsheet for further statistical treatment. As for the intervention, cephalometric analysis was performed without human supervision by the CEFBOT AI software (RadioMemory Ltd., Belo Horizonte, Brazil). Among the cephalometric analyses available in the software, Arnett's was selected, and all landmarks and measurements were performed by CEFBOT's AI (Figure 1).

To assess measurement reliability in the control (human examiner) and test (CEFBOT) groups, the radiographs were deleted from the system's memory so that they could be reevaluated by the examiner and by CEFBOT 15 days after the first records were taken. All numerical values resultant from these reevaluations were recorded in a Numbers (Apple Inc., Cupertino, Califórnia) spreadsheet for later statistical analysis.

Statistical analysis

All statistical procedures were performed using the SPSS 22.0 statistical software (SPSS Inc., Chicago, IL). The ICC was used for examiner calibration and for the assessment of measurement reliability in the control and in the test groups. The Shapiro-Wilk test

Table 1 Linear and angular cephalometric measurement reproducibility in control (human examiner) and test (CEFBOT) groups

	Control (Examiner)		*Test (Cefbot)*	
	ICC	*p* value	ICC	*p* value
DS - Glabela	0.983	<0.001[a]	Unmarked point	NP
DS - Nasal tip	0.989	<0.001[a]	0.943	<0.001[a]
DS - A'	0.958	<0.001[a]	0.997	<0.001[a]
DS – Ls	0.995	<0.001[a]	0.997	<0.001[a]
DS - 1 Sup	0.965	<0.001[a]	0.997	<0.001[a]
DS - 1 Inf	0.946	<0.001[a]	0.998	<0.001[a]
DS – Li	0.991	<0.001[a]	0.998	<0.001[a]
DS - B'	0.994	<0.001[a]	0.996	<0.001[a]
DS - Pog'	0.994	<0.001[a]	0.998	<0.001[a]
IFS – THL	0.876	<0.001[a]	0.768	<0.001[a]

DS, Distance from Subnasale; ICC, intraclass correlation coefficient.;IFP, Inclination Frankfurt plane; NP, Not performed; THL, True Horizontal Line.
Regular font, linear measumerements (millimeters). Italics, angular measurements (degrees).
[a]p ≤ 0.05 (statistically significant difference);

Table 2 Comparison of measurements from the control (human examiner) and the test group (CEFBOT)

	Control (Examiner)		*Test (CEFBOT)*		
	Mean	SD	Mean	SD	*p*-value
DS - Glabela	6.266	3.530	NP	NP	NP
DS - Nasal tip	10.490	5.115	9.922	5.024	0.511^{NS}
DS - A'	0.157	1.275	−0.031	1.248	0.767^{NS}
DS - Ls	3.063	2.366	2.246	2.266	0.197^{NS}
DS - 1 Sup	−6.211	4.681	−7.110	4.509	0.644^{NS}
DS - 1 Inf	−7.353	7.189	−9.177	5.883	0.407^{NS}
DS - Li	−1.483	3.874	−1.225	3.695	0.568^{NS}
DS - B'	−3.917	5.003	−4.220	4.348	0.935^{NS}
DS - Pog'	−1.863	4.621	−2.606	4.016	0.722^{NS}
IFP - THL	2.643	2.514	2.176	1.882	0.671^{NS}

DS, Distance from Subnasale; ICC, intraclass correlation coefficient; IFP, Inclination Frankfurt plane; NP, Not performed; NS, absence of statistically significant difference; SD, standard deviation;*p* values for the t-test for independent variablesTHL, True Horizontal Line.
Regular font, linear measumerements (millimeters). Italics, angular measurements (degrees).

was applied to check the Gaussian distribution of the data. The *t*-test for independent variables was used to assess the agreement between the two groups' means for each measurement. The level of significance was set at $p \leq 0.05$ for all tests. The classic study by Landis and Koch[10] was adopted as reference to interpret the ICC, whereby values between 0.81 and 1.00 are considered "almost perfect".

Results

An ICC value of 0.964 was obtained after examiner calibration, which indicated an excellent agreement according to Landis and Koch.[10] ICC was also applied to assess measurement reproducibility in the two groups (Table 1).

The highest ICC value for the reproducibility of measurements obtained by CEFBOT was 0.998, which is "almost perfect",[10] when considering the factors Distance from Subnasale - 1 Inf, Distance from Subnasale - L e Distance from Subnasale - Pog. On the other hand, the lowest ICC value was 0.768, still considered substantial,[10] and it was found for the factor Inclination Frankfurt Pl. - THL (Table 1). Reproducibility of the calibrated examiner, alternatively, showed the highest ICC value of 0.995 for the factor Distance do Subnasale - Ls and the lowest ICC value of 0.876 for the factor Inclination Frankfurt Pl. - THL (Table 1).

Interestingly - and particularly so, given that this is an anterior point that is not plagued by superimposition of other skeletal structures - CEFBOT was unable to measure the distance from the glabella to the subnasale point. The other results of the independent *t* test comparing the mean linear and angular cephalometric measurements between the human examiner and CEFBOT are shown in Table 2. When the agreement between the two groups was compared, no statistically significant difference was observed ($p > 0.05$)

Discussion

This study aimed to investigate whether the AI-based CEFBOT software reliably performs the Artificial Intelligence Identification (AII) of the cephalometric points pertaining to Arnett's cephalometric analysis.

Out of the 10 linear and angular measurements proposed in Arnett's analysis, CEFBOT showed almost perfect reproducibility in 8 measurements, substantial reproducibility in 1 measure and was unable to perform 1 measurement.

It is important to mention that Arnett's analysis was not originally in the list of proposed AI software analyses offered by CEFBOT. Currently, this specific analysis is unavailable for commercial use. Its inclusion on CEFBOT's plataform happened upon our request; therefore, CEFBOT had never been trained to perform it, which could explain its inability to identify the very first factor of Arnett's analysis, Distance Subnasale - Glabella'. On the other hand, CEFBOT's performance shown by our experiments was surprising at least, given that the software had not gone through machine learning sessions before our tests.

As expected, the human calibrated examiner showed almost perfect reproducibility in all but one of the measurements (Table 1). Our results differ from those obtained elsewhere regarding the reproducibility of the algorithm.[11] Interestingly, the algorithm tested in that study was superior to human landmarking in terms of reproducibility, while we found similar reproducibility for both groups in the present study.

Differences between the measurements obtained by CEFBOT and by the human examiner did not reach statistical significance, suggesting that CEFBOT shows good reproducibility for landmarking and measurement. This result ratifies those from a recent study that suggested the use of AI software as a speeding tool in cephalometric tracing - but under human supervision.[12] On the other hand, earlier studies by Leonardi et al[13] and by Shahidi et al[11] showed that, back then, the clinical application of AI-based cephalometric software was not advisable due to low reproducibility in landmarking.

Our results also differed from those of a very recent study in which the AI software landmarking reliability was reported as superior to the human one.[14] Yet, AI-based software for cephalometric tracing and measurement may be considered worthy of attention as it provides reliable values for clinical planning in less time and eliminates the subjectivity inherent to human landmarking.[14]

Among the many elements that form the notion of AI, deep learning and neural networks have gained space in automated cephalometric landmark identification.[15] While digital 3D dental exams are currently and widely used for diagnosis and planning,[13] cephalometric assessment of the dentofacial deformity patient remains a vital step for orthodontists and oral and maxillofacial surgeons.

One of the evident advantages of using an AI-based software such as CEFBOT is that it can be extremely timesaving: in this study, the identification and marking of 66 cephalometric points took less than a minute. This feat could not possibly be achieved by a human examiner, which makes CEFBOT worthy of consideration for clinical, routine use.

Another positive finding is that CEFBOT is comparable to a human examiner in terms of reproducibility. Indeed, the software showed an almost perfect agreement in 8 out of the 10 selected measurements, except for the angle between the Frankfurt horizontal plane and the true horizontal line (THL), which nevertheless showed a substantial agreement between measurements.

However, CEFBOT reliability can be considered insufficient, at least in its present version. The software failed to consistently calculate one of the measurements of Arnett's analysis, *i.e*. the glabella-subnasale distance. This measurement was calculated only in 9 out of the 30 radiographs analyzed. One hypothesis is that CEFBOT's AI has problems in the marking of the Glabella point, which is a problem solvable by means of machine learning. CEFBOT's inability to mark all the points necessary to perform Arnett's analysis autonomously highlights the fundamental role of the radiologist in cephalometric marking, and places CEFBOT not as an autonomous-intelligent system, but rather as a great example of human-machine hybrid-augmented intelligence.[16]

Apart from that isolated issue, CEFBOT seems useful to speed up Arnett's cephalometric analysis and can be used as an aid for orthodontic and surgical planning. We stress that the goal of our study is not to say that experienced professionals are easily replaceable by a machine, but rather to introduce a tool that creates, in cooperation with the radiologist, an augmented intelligence where "1 + 1>2".[16] Thus, after a lightning fast landmarking identification performed by the software, the supervising radiologist can make better use of his time by spotting incongruences and make the necessary corrections.

In short and within the limitations of methodology, our results suggest that the AI-based CEFBOT software is, at its current version, a promising tool for cephalometric point identification and marking according to Arnett's analysis - provided it is used under the supervision of a radiologist. The electronic eye still depends on the human one to be trained and perfected - and, maybe, that is for the best.

Acknowledgment

The authors are grateful to Radiomemory Ltd. (https://radiomemory.com.br/) for kindly providing the CEFBOT software for the study.

REFERENCES

1. Chen S-K, Chen Y-J, Yao C-CJ, Chang H-F. Enhanced speed and precision of measurement in a computer-assisted digital cephalometric analysis system. *Angle Orthod* 2004; **74**: 501–7. doi: https://doi.org/10.1043/0003-3219(2004)074<0501:ESAPOM>2.0.CO;2
2. Collins J, Shah A, McCarthy C, Sandler J. Comparison of measurements from photographed lateral cephalograms and scanned cephalograms. *Am J Orthod Dentofacial Orthop* 2007; **132**: 830–3. doi: https://doi.org/10.1016/j.ajodo.2007.07.008
3. Goracci C, Ferrari M. Reproducibility of measurements in tablet-assisted, PC-aided, and manual cephalometric analysis. *Angle Orthod* 2014; **84**: 437–42. doi: https://doi.org/10.2319/061513-451.1
4. Montúfar J, Romero M, Scougall-Vilchis RJ. Automatic 3-dimensional cephalometric landmarking based on active shape models in related projections. *Am J Orthod Dentofacial Orthop* 2018; **153**: 449–58. doi: https://doi.org/10.1016/j.ajodo.2017.06.028
5. Forsting M. Hot topics: will machine learning change medicine? *J Nucl Med* 2017; **58**: 357–8.
6. Hung K, Montalvao C, Tanaka R, Kawai T, Bornstein MM. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: a systematic review. *Dentomaxillofac Radiol* 2020; **49**: 20190107. doi: https://doi.org/10.1259/dmfr.20190107
7. www.radiomemory.com.br [homepage on the Internet].. Available from: http:// https://materiais.radiomemory.com.br/lp-campanha-ia [cited 2021 February 6].
8. Arnett GW, Bergman RT. Facial keys to orthodontic diagnosis and treatment planning. Part I. *American Journal of Orthodontics and Dentofacial Orthopedics* 1993; **103**: 299–312. doi: https://doi.org/10.1016/0889-5406(93)70010-L
9. Durão APR, Morosolli A, Pittayapat P, Bolstad N, Ferreira AP, Jacobs R. Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative

study. *Imaging Sci Dent* 2015; **45**: 213–20. doi: https://doi.org/10.5624/isd.2015.45.4.213

10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74. doi: https://doi.org/10.2307/2529310
11. Shahidi S, Shahidi S, Oshagh M, Gozin F, Salehi P, Danaei SM. Accuracy of computerized automatic identification of cephalometric landmarks by a designed software. *Dentomaxillofac Radiol* 2013; **42**: 20110187. doi: https://doi.org/10.1259/dmfr.20110187
12. Meriç P, Naoumova J. Web-Based fully automated cephalometric analysis: comparisons between App-aided, computerized, and manual tracings. *Turk J Orthod* 2020; **33**: 142–9. doi: https://doi.org/10.5152/TurkJOrthod.2020.20062
13. Leonardi R, Giordano D, Maiorana F. An evaluation of cellular neural networks for the automatic identification of cephalometric landmarks on digital images. *J Biomed Biotechnol* 2009; **2009**: 1–12. doi: https://doi.org/10.1155/2009/717102
14. Hwang H-W, Park J-H, Moon J-H, Yu Y, Kim H, Her S-B, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? *Angle Orthod* 2020; **90**: 69–76. doi: https://doi.org/10.2319/022019-129.1
15. Sam A, Currie K, Oh H, Flores-Mir C, Lagravére-Vich M. Reliability of different three-dimensional cephalometric landmarks in cone-beam computed tomography : A systematic review. *Angle Orthod* 2019; **89**: 317–32. doi: https://doi.org/10.2319/042018-302.1
16. Pan Y. Heading toward artificial intelligence 2.0, engineering. 2016; **2**: 409–13.