Research, Society and Development, v. 11, n. 14, e73111435820, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i14.35820

Avaliação da performance de dois softwares com inteligência artificial por meio das medidas geradas pela análise de Mcnamara em telerradiografia cefalométrica lateral

Evaluation of the performance of two software artificial intelligence-based by means of the measurements according to Mcnamara's Analysis in lateral cephalometric radiographs

Evaluación del rendimiento de dos softwares con inteligencia artificial mediante las medidas generadas por el análisis de Mcnamara en radiografías cefalométricas laterales

Recebido: 26/09/2022 | Revisado: 12/10/2022 | Aceitado: 15/10/2022 | Publicado: 19/10/2022

Laura Luiza Trindade de Souza
ORCID: https://orcid.org/0000-0002-8956-9708
Universidade Federal de Sergipe, Brasil
E-mail: latrindasouza2198@gmail.com
Thaísa Pinheiro Silva
ORCID: https://orcid.org/0000-0002-7485-0206
Universidade Estadual de Campinas, Brasil
E-mail: thaisapinheirosilva@hotmail.com
William José e Silva Filho
ORCID: https://orcid.org/0000-0002-2117-3352
Universidade de Pernambuco, Brasil
E-mail: williamfilho10@hotmail.com
Bruno Natan Santana Lima
ORCID: https://orcid.org/0000-0003-2828-2129
Universidade Federal de Sergipe, Brasil
E-mail: brunonatanufs@gmail.com
Amanda Caroline Nascimento Meireles
ORCID: https://orcid.org/0000-0001-6007-9257
Universidade Federal de Sergipe, Brasil
E-mail: meireles.carolinie@gmail.com
Iris Tamara de Santana Oliveira
ORCID: https://orcid.org/0000-0002-7463-591X
Universidade Federal de Sergipe, Brasil
E-mail: iristamara38@gmail.com
Wilton Mitsunari Takeshita
ORCID: https://orcid.org/0000-0001-5682-1498
Universidade Estadual Paulista, Brasil
E-mail: wmtakeshita2@gmail.com

Resumo

O objetivo do trabalho foi comparar a performance de dois softwares com IA em telerradiografia cefalométrica lateral, por meio da avaliação da reprodutibilidade e confiabilidade das medidas lineares e angulares da análise de McNamara. Foram marcadas 30 telerradiografias cefalométricas por meio do método digital pelo examinador no Radiocef (RadioMemory). Posteriormente, a amostra foi marcada por meio da IA dos softwares CEFBOT (RadioMemory) e WebCeph™ (AssembleCircle), para avaliação da reprodutibilidade e confiabilidade, em relação ao examinador e os softwares em questão. Para calibrar o examinador e avaliar a confiabilidade das marcações do examinador, CEFBOT, e WebCeph™ utilizou o Coeficiente de Correlação Intraclasse (ICC), bem como, o teste ANOVA e pós teste de Tukey avaliou a reprodutibilidade dos softwares, por meio dos pontos cefalométricos que compõem a análise de McNamara. O ICC médio do examinador, CEFBOT e do WebCeph foram 0.960, 0.940 e 0.954, respectivamente, indicando concordância quase perfeita. Ao comparar CEFBOT com examinador, observou-se diferença estatística ($p<0.01$) apenas na medida A-N perpendicular. Quanto ao WebCeph™, ao comparar com o examinador houve diferença significativa entre os fatores dois ao seis e o dez. E comparado ao CEFBOT, houve divergência nos mesmos fatores somado ao fator onze. Além disso, o WebCeph™ não identificou as medidas Nfa-Nfp e Bfa-Bfp. O CEFBOT apresentou reprodutibilidade e confiabilidade na identificação dos pontos cefalométricos determinados pela análise de McNamara, mas necessitando de supervisão humana. O WebCeph apresentou concordância quase perfeita nas marcações, porém seis medidas apresentaram-se diferentes do examinador e duas não foram realizadas pela aplicação.

Palavras-chave: Inteligência artificial; Ortodontia; Aprendizado de máquina; Radiologia; Diagnóstico.

Abstract
The aim of this study was to compare the performance of two software programs with AI in lateral cephalometric teleradiography by assessing the reproducibility and reliability of the linear and angular measurements of McNamara's analysis. Thirty cephalometric teleradiographs were marked using the digital method by the examiner in Radiocef (RadioMemory). Subsequently, the sample was marked using the CEFBOT (RadioMemory) and WebCeph™ (AssembleCircle) software AI to evaluate the reproducibility and reliability of the examiner and the software. To calibrate the examiner and evaluate the reliability of the examiner, CEFBOT, and WebCeph™ markings, the Intraclass Correlation Coefficient (ICC) was used, as well as the ANOVA test and Tukey's post-test evaluated the reproducibility of the software, using the cephalometric landmarks that comprise McNamara's analysis. The mean ICC of the examiner, CEFBOT and WebCeph were 0.960, 0.940 and 0.954, respectively, indicating almost perfect agreement. When comparing CEFBOT with examiner, statistical difference ($p<0.01$) was observed only in the perpendicular A-N measurement. As for WebCeph™, when comparing with the examiner there was a significant difference between factors two to six and ten. And compared to CEFBOT, there was divergence in the same factors plus factor eleven. In addition, WebCeph™ did not identify the measurements Nfa-Nfp and Bfa-Bfp. CEFBOT showed reproducibility and reliability in identifying the cephalometric landmarks determined by McNamara's analysis but required human supervision. WebCeph showed almost perfect agreement in the markings, but six measurements were different from the examiner and two were not performed by the application.
Keywords: Artificial intelligence; Orthodontics; Machine learning; Radiology; Diagnosis.

Resumen
El objetivo de este estudio fue comparar el rendimiento de dos programas informáticos con IA en la telerradiografía cefalométrica lateral, evaluando la reproducibilidad y la fiabilidad de las medidas lineales y angulares del análisis de McNamara. Treinta telerradiografías cefalométricas fueron marcadas mediante el método digital por el examinador en Radiocef (RadioMemory). Posteriormente, la muestra se marcó utilizando la IA del software CEFBOT (RadioMemory) y WebCeph™ (AssembleCircle) para evaluar la reproducibilidad y la fiabilidad en relación con el examinador y el software en cuestión. Para calibrar el examinador y evaluar la fiabilidad de las marcas del examinador, del CEFBOT y del WebCeph™, se utilizó el coeficiente de correlación intraclase (CCI), así como la prueba ANOVA y la prueba posterior de Tukey evaluaron la reproducibilidad del software, utilizando los puntos de referencia cefalométricos que componen el análisis de McNamara. El CCI medio del examinador, del CEFBOT y del WebCeph fue de 0,960, 0,940 y 0,954, respectivamente, lo que indica una concordancia casi perfecta. Al comparar el CEFBOT con el examinador, se observaron diferencias estadísticas ($p<0,01$) sólo en la medición perpendicular A-N. Al comparar WebCeph™ con el examinador, se observó una diferencia significativa entre los factores dos a seis y diez. En comparación con el CEFBOT, hubo divergencia en los mismos factores más el factor once. Además, WebCeph™ no identificó las medidas Nfa-Nfp y Bfa-Bfp. El CEFBOT mostró reproducibilidad y fiabilidad en la identificación de los puntos de referencia cefalométricos determinados por el análisis de McNamara, pero requirió supervisión humana. WebCeph mostró una concordancia casi perfecta en las marcas, pero seis mediciones fueron diferentes a las del examinador y dos no fueron realizadas por la aplicación.
Palabras clave: Inteligência artificial; Ortodoncia; Aprendizaje automático; Radiología; Detección.

1. Introdução

A telerradiografia cefalométrica lateral é um exame complementar utilizado para o diagnóstico e tratamento de discrepâncias esqueléticas, relações dentárias e de tecidos moles das estruturas faciais (Albarakati et al., 2012; Debelmas et al., 2019). A partir da introdução do cefalostato, por Broadbent, em 1931, foi possível a padronização do posicionamento da cabeça do paciente durante a execução da técnica radiográfica, em consequência disso, o exame tornou-se uma ferramenta essencial no planejamento ortodôntico e cirúrgico (Chien et al., 2009; Olmez et al., 2011; Livas et al., 2019; Khan et al., 2020; Meriç & Naoumova, 2020).

Antigamente, os traçados cefalométricos somente eram realizados manualmente, sendo cada ponto cefalométrico identificado de forma manual em um papel acetato. Contudo, o método manual demanda tempo e há uma tendência à subjetividade do profissional que irá realizar os traçados, o que pode levar a grandes discrepâncias no diagnóstico e planejamento ortodôntico e/ou cirúrgico. Com isso, o método digital foi inserido e as telerradiografias cefalométricas laterais passaram a ser digitalizadas e o traçado realizado em softwares de telerradiografia cefalométrica lateral (Chen et al., 2004; Albarakati et al., 2012).

Com o crescimento exponencial da tecnologia e ciência, a realização da análise cefalométrica de forma digital passou

a substituir o método convencional (Albarakati et al., 2012). Com isso, surgiram programas computadorizados de análise cefalométrica, que quando comparados ao método manual, possuem benefícios como possibilidade de compartilhamento e aprimoramento de imagens, identificação de pontos cefalométricos em menor tempo e automatização de etapas como desenho do cefalograma e geração dos valores lineares e angulares (Meriç & Naoumova, 2020; Khanagar et al., 2021). No entanto, esses programas ainda necessitam que o profissional localize manualmente os pontos cefalométricos, e, portanto, a subjetividade na identificação dos pontos permanece inerente ao método (Hung et al., 2019).

Assim, na tentativa de otimizar o tempo do profissional e reduzir a subjetividade inerente a essa tarefa, há um crescente desenvolvimento de softwares com inteligência artificial com a finalidade de identificação dos pontos cefalométricos. E a inteligência artificial (IA) pode ser definida como uma constelação de itens (algoritmos, robótica, redes neurais) que permitem a um software propriedades de inteligência comparáveis as de um ser humano, dentre elas, o aprendizado de bancos de dados com mínima interferência humana (Forsting, 2017), sendo cada vez mais difundida e imperativa na resolução de problemas complexos e ainda em usos mais triviais que ultrapassam fronteiras como um simples software de tradução ou assistente virtual (Obermeyer & Emanuel, 2016; Chen et al., 2020).

Dentre as possibilidades de uso das análises, a análise de McNamara é bastante referenciada em ser a primeira análise que não é apenas sensível à posição dos dentes, mas também às estruturas da base do crânio (Mcnamara, 1984; Ravikumar et al., 2019). Devido aos riscos que os erros na avaliação cefalométrica podem implicar, um deles é o tratamento inadequado. Diversos estudos surgiram com o intuito de analisar a reprodutibilidade das marcações automatizadas com IA, e assim, aumentar a qualidade e precisão do tratamento ortodôntico (Farooq, 2016; Khanagar et al., 2021).

Na Odontologia, a IA vem sendo aplicada em diferentes vertentes, e uma delas é na análise de exames digitais para diagnóstico e planejamento. Para isso, foram desenvolvidos softwares que, por meio da análise de um grande volume de dados, equipados com algoritmos aperfeiçoados por especialistas, conseguem identificar e reproduzir padrões (Hung et al., 2019; Kunz et al., 2019). Para avaliar a performance de um software, estuda-se a sua a reprodutibilidade, isto é, a concordância entre a mesma medida marcadas por dois métodos distintos e a confiabilidade, ou seja, a concordância entre uma mesma medida marcada pelo mesmo método em dois momentos diferentes (Zamrik & Iseri, 2021).

A Ortodontia tornou-se uma das mais importantes áreas de aplicação da IA (Chen et al., 2020). Diversos softwares com IA já estão disponíveis no mercado, todavia, a maioria deles possuem licença paga, o que torna menos acessível para os profissionais utilizarem na prática clínica. Diante do exposto, a pesquisa tem como objetivo comparar a performance de dois softwares com IA — WebCeph™ (AssembleCircle), disponível gratuitamente, e CEFBOT (RadioMemory Ltd., Belo Horizonte, Brasil), disponível de forma paga— por meio da avaliação da confiabilidade e reprodutibilidade da marcação cefalométrica automatizada na análise de McNamara (1984).

2. Metodologia

2.1 Princípios éticos e cálculo amostral

Trata-se de um estudo observacional, retrospectivo e comparativo. Este estudo foi cadastrado com (CAAE: 47835221.5.0000.5546) e aprovado pelo Comitê de Ética e Pesquisa em seres humanos sob o parecer 4.919.920, na Universidade Federal de Sergipe (UFS), Hospital Universitário, estando de acordo com a declaração de Helsinki e realizado conforme a iniciativa STROBE. Devido à natureza retrospectiva deste estudo, o consentimento informado assinado não foi exigido pelo Comitê.

Foi realizado cálculo amostral em que o coeficiente de correlação intraclasse (ICC) foi de 0.70, com poder de teste de 99% e nível de significância de 5%, para tanto, foram necessárias 28 telerradiografias cefalométricas laterais, baseado nos estudos de Durão et al. (2015) e Silva et al. (2021).

2.2 Características da amostra

Foram selecionadas trinta telerradiografias cefalométricas laterais digitais do arquivo da disciplina de Radiologia, no ambulatório de Radiologia, localizado no Hospital Universitário da Universidade Federal de Sergipe. O critério de inclusão contou com telerradiografias cefalométricas em norma lateral de pacientes sem distinção de sexo ou idade. E como critérios de exclusão foram considerados: (1) mau posicionamento da cabeça no cefalostato; (2) pacientes com deformidades craniofaciais graves e assimetrias faciais; (3) pacientes com pinos, placas ou parafusos sobrepondo estruturas anatômicas de interesse.

2.3 Estudo em diferentes softwares de inteligência artificial (CEFBOT e WEBCEPH)

Após seleção da amostra, as telerradiografias cefalométricas laterais foram salvas nos formatos de imagem JPEG. Sendo então armazenadas e analisadas em um computador de uso pessoal Inspiron 7472 14 polegadas (Dell, Intel Core i5, 8 GB RAM, armazenamento 1TB (5.4K), tela Full HD). Utilizou-se a resolução de 600dpi Bissoli et al. (2007).

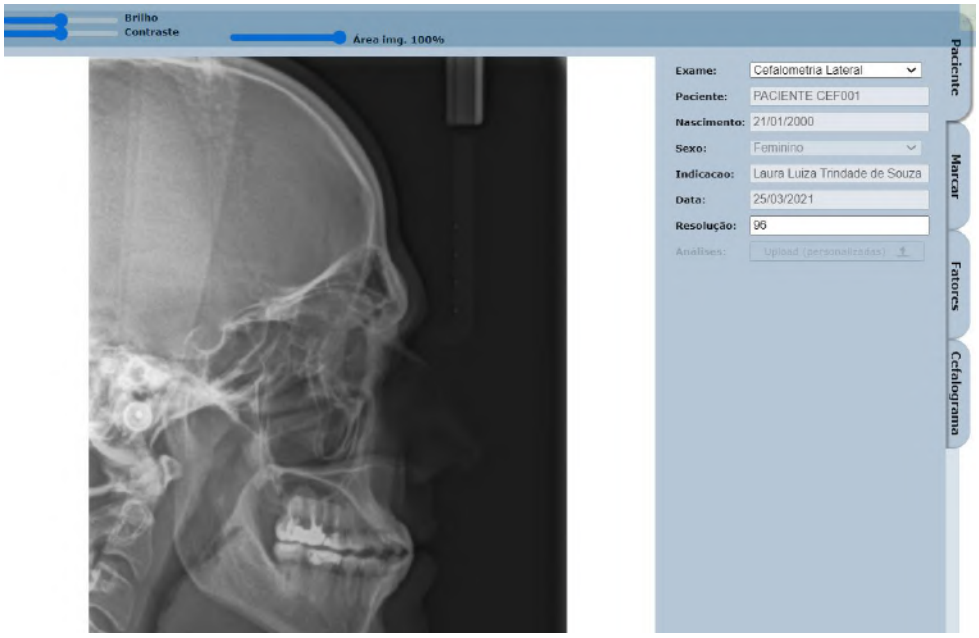
Para início das análises cefalométricas foi necessária a calibração do examinador, e para tanto, 30% da amostra foi utilizada. O examinador foi treinado previamente por radiologista com mais 20 anos de experiência em traçados cefalométricos computadorizados, e, somente quando o Coeficiente de Correlação Intraclasse (ICC) foi superior a 0.90, o traçado pelo examinador calibrado foi iniciado.

Para que as marcações pudessem ser realizadas, solicitou-se uma licença de uso do software CEFBOT, da empresa RadioMemory e foi realizado cadastro na plataforma WebCeph™, da empresa Assemble Cicle. Assim, foram selecionadas, de acordo com os critérios de inclusão e exclusão, 30 telerradiografias cefalométricas laterais que foram marcadas através do método digital pelo examinador calibrado utilizando a própria plataforma do software Radiocef (RadioMemory). Posteriormente, as mesmas telerradiografias cefalométricas laterais foram marcadas através da IA dos softwares CEFBOT (RadioMemory) e WebCeph™ (AssembleCircle), para avaliação de duas vertentes de marcação: reprodutibilidade e confiabilidade em relação ao examinador e os softwares em questão.

A marcação foi realizada por um examinador devidamente calibrado. O examinador calibrado foi responsável pela identificação dos pontos cefalométricos do grupo controle, com base na análise McNamara (Mcnamara, 1984) de tal forma a serem geradas grandezas lineares e angulares na plataforma do Radiocef (RadioMemory).

A mesma amostra foi submetida à marcação automatizada apor meio da IA do software CEFBOT (RadioMemory) (Figura 1), e por meio da IA do software WebCeph™ (AssembleCircle) (Figura 2). As grandezas lineares e angulares geradas pelos sistemas através da IA foram armazenadas em planilha no Office Excel 2019 para posterior análise.

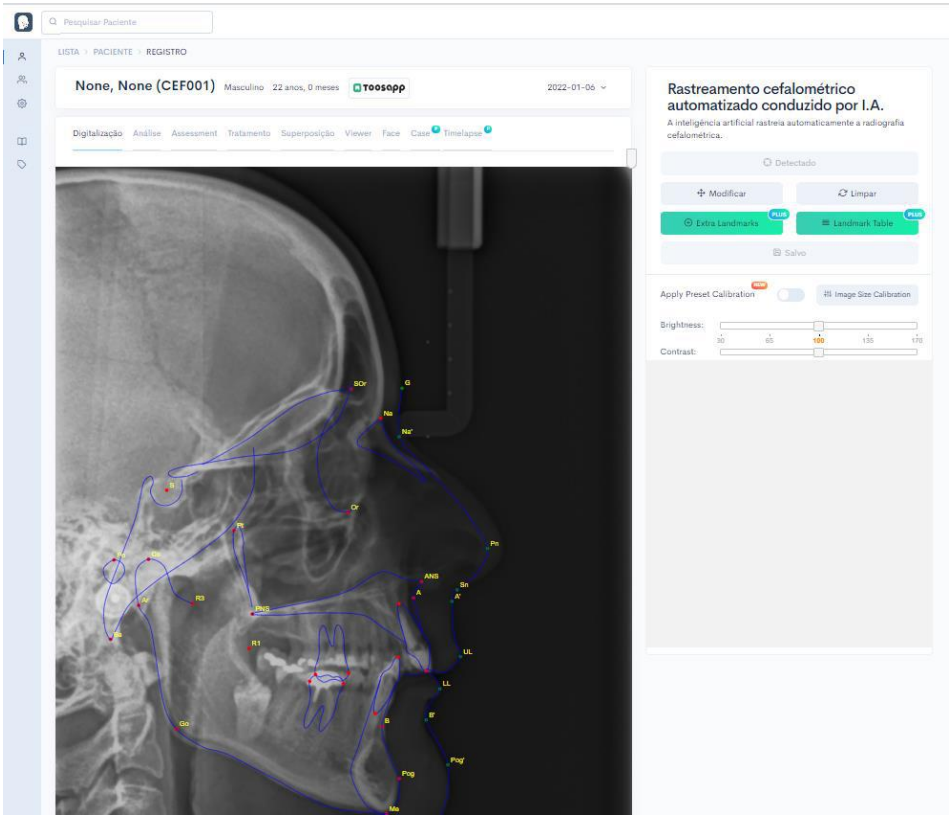
Figura 1. Interface do software CEFBOT



Fonte: Site Radiomemory.

Na Figura 1 pode-se visualizar a interface do software CEFBOT, um sistema baseado em machine learning (ML) que estima a posição de pontos pré-definidos (pontos anatómicos) a partir da imagem de uma radiografia.

Figura 2. Interface do software WebCeph™ (AssembleCircle).



Fonte: Site WebCeph™

Na Figura 2 ilustra a o software WebCeph, uma plataforma baseada em inteligência artificial, intuitiva e que permite ao profissional realizar a marcação automatizada, como também, organizar toda a documentação do paciente.

Para a avaliação da confiabilidade, após 15 dias do primeiro registro, foi realizada nova marcação pelos examinadores humanos e pelos dois softwares de IA seguindo o mesmo padrão metodológico. Para os dois softwares, os registros anteriormente realizados foram deletados da memória do sistema e, em seguida, as radiografias foram novamente "upadas" e submetidas à marcação automática pela IA do software. Os valores obtidos na primeira e segunda marcação de todos os grupos foram armazenados em planilha no Office Excel 2019 para posterior análise.

2.4 Análise estatística

Os procedimentos estatísticos foram realizados no programa estatístico BioEstat 5.3 (Instituto Mamirauá, Belém, Brasil). A calibração do examinador foi realizada pelo Coeficiente de Correlação Intraclasse (ICC). A confiabilidade da medição executada pelos examinadores e os programas CEFBOT (RadioMemory) e WebCeph™ (AssembleCircle), foram calculadas pelo coeficiente de correlações intraclasses (ICC) entre a primeira marcação e a segunda marcação. Com a finalidade de investigar a reprodutibilidade entre os grupos utilizou o teste ANOVA e o teste de Tukey para variáveis independentes. O nível de significância foi estabelecido em $p<0.05$, para todos os testes realizados.

3. Resultados e Discussão

O teste de calibração avaliado pelo coeficiente de correlação Intraclasse (ICC) entre o examinador após treinamento e o examinador experiente obteve valor maior que 0.9, correspondente à concordância excelente, de acordo com Landis (1997).

A Tabela 1 apresenta os resultados da avaliação da reprodutibilidade dos softwares com o examinador calibrado. Quando comparado o CEFBOT (RadioMemory) com o examinador calibrado, observou-se diferença estatística ($p<0.01$) apenas no fator um (A-N perpendicular).

Quanto ao WebCeph™ (AssembleCircle), ao comparar com o examinador houve diferença significativa entre os fatores dois ao seis e o dez. E comparado ao CEFBOT (RadioMemory), houve divergência nos mesmos fatores somado ao fator onze. Além disso, o WebCeph™ (AssembleCircle) não conseguiu marcar os fatores doze (Nfa-Nfp) e treze (Bfa-Bfp) (Tabela 1).

O ICC também foi aplicado para avaliar a confiabilidade das medições nos três grupos (Tabela 2). Ao avaliar a confiabilidade dos diferentes métodos de avaliação, o maior valor de ICC foi identificado no CEFBOT (RadioMemory), apresentando 1.000 para o fator Prn.(Sn-Ls), o que é "quase perfeito" (Landis & Koch, 1977). Por outro lado, o valor de ICC mais baixo foi 0.611, relacionado ao terceiro fator da análise (Co-Gn), todavia, ainda considerado de concordância substancial (Landis & Koch, 1977). Ainda a respeito do CEFBOT (RadioMemory), com exceção do terceiro fator, todos os outros fatores, apresentaram valores considerados "quase perfeito" (Landis & Koch, 1977). No que tange a confiabilidade do WebCeph™ (AssembleCircle), pôde-se observar o valor de ICC mais alto para os fatores nove (Pog-N perpendicular) e dez (SF1/-A perpendicular) da análise, apresentando 0.999, e o menor valor para o primeiro fator (A-N perpendicular), apresentando 0.781. Os fatores doze (Nfa-Nfp) e treze (Bfa-Bfp) não puderam ser avaliados uma vez que não foram gerados pelo WebCeph™ (AssembleCircle) (Tabela 2).

Em relação à confiabilidade do examinador calibrado, o maior valor de ICC foi de 0.991 para o terceiro e o sexto fator, e o menor valor foi de 0.894 para o quinto fator (Tabela 2). Por fim, o ICC médio gerado pelas marcações cefalométrica do examinador calibrado, do CEFBOT e do WebCeph foram 0.960, 0.940 e 0.954, respectivamente, indicando concordância quase perfeita.

Tabela 1. Valores de média, Desvio Padrão (DP) e teste ANOVA com pós teste de Tukey para variáveis independentes entre o examinador, o programa CEFBOT e WebCeph para análise Mcnamara.

	Examinador		CEFBOT		WebCeph		E x C	E x W	C x W
	Média	DP	Média	DP	Média	DP	Valor de p	Valor de p	Valor de p
1. A-N perpendicular	1.09	8.36	0.27	9.06	-0.05	3.46	NS	NS	NS
2. *Prn.(Sn-Ls)*	102.35	12.77	114.06	11.16	81.31	4.42	<0.01*	<0.01*	<0.01*
3. Co-Gn	227.11	18.19	223.53	25.51	112.74	8.30	NS	<0.01*	<0.01*
4. Co-A	170.94	9.24	169.76	9.90	82.50	4.66	NS	<0.01*	<0.01*
5. Diferença Mx - Md	55.36	14.63	56.78	14.55	30.24	5.88	NS	<0.01*	<0.01*
6. Ena-Me	129.40	15.04	130.53	14.86	67.98	6.33	NS	<0.01*	<0.01*
7. *(Po-Or).(Go-Me)*	24.94	5.73	25.77	5.78	24.41	5.61	NS	NS	NS
8. *(Ba-N).(Ptm-Gn)*	-1.42	4.52	-2.12	4.01	-2.49	3.84	NS	NS	NS
9. Pog-N perpendicular	0.39	16.82	-1.55	15.43	-5.07	5.97	NS	NS	NS
10. SF1/-A perpendicular	16.06	8.28	15.24	7.87	4.93	3.21	NS	<0.01*	<0.01*
11. lii-(A-Pog)	6.36	7.83	7.12	7.21	2.62	3.29	NS	NS	<0.01*
12. Nfa-Nfp	27.44	7.72	28.71	8.54	NI	NI	NS	NI	NI
13. Bfa-Bfp	19.64	6.97	22.72	6.35	NI	NI	NS	NI	NI

Fontes regulares, medições lineares (milímetros); itálico, medições angulares (graus); DP: desvio padrão; NI: Não Identificado; NS: diferença não significativa. *Diferença estatisticamente significativa (p<0.01). E: examinador; C: CEFBOT; W: WebCeph. Fonte: Autores.

Tabela 2. Medidas cefalométricas com grandeza linear e angular aplicando o Índice de Correlação Inter-classe (ICC) para análise de McNamara.

	Examinador		CEFBOT		WebCeph	
	ICC	Valor de p	ICC	Valor de p	ICC	Valor de p
1. A-N perpendicular	0.989	<0.001*	0.937	<0.001*	0.781	<0.001*
2. *Prn.(Sn-Ls)*	0.957	<0.001*	1.000	<0.001*	0.996	<0.001*
3. Co-Gn	0.991	<0.001*	0.611	<0.001*	0.950	<0.001*
4. Co-A	0.945	<0.001*	0.961	<0.001*	0.916	<0.001*
5. Diferença Mx - Md	0.894	<0.001*	0.985	<0.001*	0.983	<0.001*
6. Ena-Me	0.991	<0.001*	0.999	<0.001*	0.970	<0.001*
7. *(Po-Or).(Go-Me)*	0.983	<0.001*	0.987	<0.001*	0.997	<0.001*
8. *(Ba-N).(Ptm-Gn)*	0.963	<0.001*	0.896	<0.001*	0.999	<0.001*
9. Pog-N perpendicular	0.967	<0.001*	0.954	<0.001*	0.999	<0.001*
10. SF1/-A perpendicular	0.990	<0.001*	0.956	<0.001*	0.926	<0.001*
11. lii-(A-Pog)	0.960	<0.001*	0.971	<0.001*	0.981	<0.001*
12. Nfa-Nfp	0.930	<0.001*	0.974	<0.001*	PNI	NI
13. Bfa-Bfp	0.921	<0.001*	0.994	<0.001*	PNI	NI

Fontes regulares, medições lineares (milímetros); itálico, medições angulares (graus). *Diferença estatisticamente significativa (p<0.05). PNI: Ponto Não Identificado; NI: Não Identificado. Fonte: Autores.

A presente pesquisa teve como objetivo comparar a performance dos softwares com inteligência artificial, CEFBOT (RadioMemory) e WebCeph™ (AssembleCircle), de acordo com a avaliação da reprodutibilidade e confiabilidade, das medidas lineares e angulares da análise de McNamara.

Na Ortodontia, as principais considerações para um tratamento ortodôntico eficaz são: um diagnóstico preciso, um excepcional plano de tratamento e um bom prognóstico (Subramanian *et al.*, 2022). Dado que muitos tratamentos são irreversíveis ou causam efeitos secundários permanentes, tais como: reabsorção radicular e recessão gengival. Por essa razão, é importante ter ferramentas que aumentem a precisão do tratamento proposto (Yu *et al.*, 2020).

Em nosso estudo foram avaliadas as medidas lineares e angulares ao invés da localização dos pontos cefalométricos, uma vez que, as medições são o produto do processo da identificação cefalométrica e fornecem dados para o planejamento e tratamento ortodôntico (Ongkosuwito *et al.*, 2002; Santoro et al., 2006; Shahidi *et al.*, 2013).

Ao comparar a performance dos dois softwares com o examinador devidamente calibrado foi avaliado a performance de cada um. Assim, dos 13 fatores presentes na análise de McNamara, quando comparado o CEFBOT (RadioMemory) com o examinador, observou-se diferença estatística ($p<0.01$) apenas na medida Prn.(Sn-Ls). O que demonstra excelente capacidade do software marcar corretamente os pontos cefalométricos da análise em questão.

É importante mencionar o recente estudo de Silva *et al*. (2021), de metodologia semelhante ao nosso, o qual buscou avaliar o CEFBOT (RadioMemory). Em relação à reprodutibilidade, embora o software não tenha gerado um fator da análise de Arnett, os outros fatores não tiveram diferença significativa comparado com o examinador calibrado. Baseado nisso, pode-se perceber que o software é reprodutível, além disso, está em constante evolução, dado que no nosso estudo todos os fatores foram gerados e com excelente resultado.

Em relação à reprodutibilidade do WebCeph™ (AssembleCircle), quando comparado com o examinador houve diferença significativa entre os fatores 2, 3, 4, 5, 6 e o 10. E ao comparar com o CEFBOT (RadioMemory), houve divergência nos mesmos fatores e no 11. Além disso, o WebCeph™ (AssembleCircle) não conseguiu marcar os fatores 12 e 13 (Tabela 1). E essas diferenças são consideradas significativas para a precisão do diagnóstico.

O estudo de Moon *et al*. (2020) investigou qual o número necessário de dados para a aprendizagem de um software baseado em aprendizado profundo da máquina. Desse modo, concluíram que a acurácia da IA é diretamente proporcional à quantidade de dados disponível no banco de aprendizagem e ao número de pontos que o software consegue identificar. Baseado nisso, podemos afirmar que os fatores não gerados pelo WebCeph™, assim como, os quais não foram considerados confiáveis comparados ao examinador e ao CEFBOT podem ser aprimorados com a ampliação do banco de dados e aprendizagem de máquina.

E para avaliar a capacidade dos softwares marcarem diversas vezes no mesmo local foi avaliada a confiabilidade. Sendo assim, cada software marcou duas vezes a mesma telerradiografia cefalométrica lateral, e mostraram uma concordância quase perfeita em 12 medições e uma concordância substancial em 1 medida (Tabela 2).

Para avaliar a confiabilidade dos dois softwares analisou-se a marcação realizada por um examinador calibrado mostrando uma concordância quase perfeita em todas as medidas (Tabela 2). E nos três casos (examinador, CEFBOT e WebCeph™) foi possível identificar pontos em locais idênticos aos identificados anteriormente, formando-se grandezas lineares e angulares sem diferença significativa como visto na Tabela 2. Logo, é possível afirmar que, em termos de confiabilidade, os dois softwares são equiparáveis ao examinador.

Os nossos resultados em relação à confiabilidade do WebCeph™ foram semelhantes à recente pesquisa de Mahto *et al.* (2022). Sendo necessário destacar que, apesar do objetivo ter sido avaliar a reprodutibilidade e confiabilidade do software, na verdade, o estudo apenas utilizou o teste estatístico de Índice de Correlação Inter-classe (ICC). Isto é, comparou a capacidade do software identificar diversas vezes o mesmo local. E ao avaliar somente com esse teste não é possível afirmar se os pontos identificados pelo programa estão corretos. E esse fato é importante ser avaliado, dado que, em nosso estudo houve discordância em relação à sua reprodutibilidade, em razão do software ter tido divergência em vários fatores comparado ao CEFBOT (RadioMemory) e ao examinador.

Ao passo que, estudos mais antigos como os de Leonardi, Giordano & Maiorana (2009) e Shahidi *et al. (*2013) demostraram que a aplicação clínica de softwares em IA ainda não seria viável, visto que os resultados não mostraram muita reprodutibilidade na marcação dos pontos. Nossos resultados também diferem dos de Hwang *et al.* (2020) no qual o software de IA alcançou uma reprodutibilidade considerada superior à marcação humana. O que podemos afirmar, a partir do nosso estudo, é que os softwares podem complementar na rotina do profissional, mas sem substitui-lo.

Outras pesquisas como a de Kunz *et al*. (2019) e Meriç e Naoumova (2020) puderam concluir que uso de software baseado em IA é uma importante ferramenta auxiliar para otimizar a identificação cefalométrica; ainda que houve diferenças estatísticas em algumas medidas e ambos apresentaram no ângulo SN.GoGn. Curiosamente, dentre os pontos que formam esse ângulo é o SN, esse mesmo ponto faz parte do fator que apresentou o menor ICC do CEFBOT (RadioMemory) e pode-se levantar a hipótese de uma maior dificuldade de o software marcar esse ponto por ser em tecido mole.

A Ortodontia é uma especialidade que continuará a evoluir, especialmente com os avanços da IA (Masse, 2019). E dentre as vantagens dos softwares de IA, pode-se destacar o fato que uma marcação cefalométrica, realizada manualmente, demanda um tempo médio de 15 a 20 minutos (Dreyer & Raymond Geis, 2017). Por consequência, as utilizações de softwares baseados em IA, como o CEFBOT e WebCeph™, são uma excelente ferramenta para otimizar o tempo (Silva *et al.*, 2021). Neste estudo, a identificação e marcação de 59 pontos cefalométricos demorou menos de um minuto. E esse marco não poderia ser alcançado através de um examinador humano, o que torna o uso de softwares baseado em IA relevante para prática clínica.

Outro fato positivo é que o CEFBOT (RadioMemory) é comparável a um examinador humano em termos de reprodutibilidade e confiabilidade. De fato, o software mostrou uma concordância quase perfeita em 12 dos 13 fatores em termo de confiabilidade e apenas 1 fator apresentou reprodutibilidade com diferença estatística comparado ao examinador padrão ouro.

No entanto, a reprodutibilidade do WebCeph™ pode ser considerada insuficiente, pelo menos na sua versão atual. O software não conseguiu calcular as duas medidas que avaliam as vias aéreas da análise de McNamara (os espaços nasofaríngeo e bucofaríngeo). Além disso, apresentou diferença estatística em seis fatores comparado ao examinador e ao CEFBOT (RadioMemory).

Como limitações do nosso estudo, podemos apontar que a pesquisa não avaliou a possibilidade de interferência na precisão de marcação na presença de fatores metálicos como aparelhos ortodônticos, grades cirúrgicas, dentre outros, apesar de não ser ter sido o objetivo do estudo. Assim, novos estudos são sugeridos para elucidar realmente essas questões referentes à presença de fatores, que podem contribuir para diminuição da confiabilidade dos softwares em questão.

4. Conclusão

A avaliação da reprodutibilidade e confiabilidade das marcações realizadas pelo CEFBOT (RadioMemory) por meio da análise de McNamara mostrou que o software é reprodutível visto que não houve diferenças significativas entre a marcação do software e a marcação do examinador calibrado. A confiabilidade do software, assim como a do examinador calibrado, mostrou que a segunda marcação foi realizada em posição idêntica à marcação anterior, e, portanto, de excelente confiabilidade.

Em relação ao WebCeph™ (AssembleCircle), o software também apresentou uma excelente confiabilidade. No entanto, no que diz respeito à reprodutibilidade o programa apresentou diferença significava e, assim, não conseguiu marcar alguns fatores corretamente nem identificar dois fatores da análise de McNamara. Como alternativa, o profissional pode utilizar a opção de correção manual de pontos de referência fornecida pelo "WebCeph"™ e, assim, pode melhorar a precisão das medidas cefalométricas.

Da perspectiva clínica, o objetivo desse estudo não é substituir o profissional experiente, mas sim, suplementar,

colaborar, otimizar e ampliar a performance do profissional, e, portanto, a partir desta pesquisa foram avaliadas a reprodutibilidade e confiabilidade dos softwares CEFBOT (RadioMemory) e WebCeph™ (AssembleCircle), demonstrando, por fim, a sua capacidade de somar no ambiente clínico odontológico.

Por fim, considerando o crescente desenvolvimento de softwares baseados em IA, sugere-se a realização de novos estudos na área que possibilitem ampliar a performance desses programas. E para isso, é necessário desenvolver pesquisas com intuito de avaliar a reprodutibilidade e confiabilidade de softwares utilizando diferentes variáveis, a exemplo: diferentes tipos faciais, contraste de brilho, diferentes extensões de imagem, entre outros.

Referências

Albarakati, S., Kula, K., & Ghoneima (2012). A. The reliability and reproducibility of cephalometric measurements: a comparison of conventional and digital methods. *Dentomaxillofacial Radiology*, 41 (1), 11–17.

Bissoli, C. F., Takeshita, W.M., Castilho, J.C.M., Médici-Filho, E.M (2007). Digitalização de imagens em radiologia: uma nova visão de futuro. *Revista Odonto*, 30 (15), 34-39.

Borba, A. M.; Haupt, D.; Almeida Romualdo, L. T. De; Silva, A. L. F. Da; Graça Naclério-Homem, M. Da; Miloro, M (2016). How Many Oral and Maxillofacial Surgeons Does It Take to Perform Virtual Orthognathic Surgical Planning? *Journal of Oral and Maxillofacial Surgery* 74 (9), 1807–1826.

Chen, S.-K., Chen, Y.-J., Yao, C.-C. J., Chang, H.-F (2004). Enhanced Speed and Precision of Measurement in a Computer-Assisted Digital Cephalometric Analysis System. *Angle Orthodontist,*74 (4), 1-11.

Chen, Y., Stanley, K., & Att, W (2020). Artificial intelligence in dentistry: current applications and future perspectives. *Quintessence International,* 51 (3), 248–257.

Chien, P., Parks, E., Eraso, F., Hartsfield, J., Roberts, W., et al. (2009). Comparison of reliability in anatomical landmark identification using two-dimensional digital cephalometrics and three-dimensional cone beam computed tomography in vivo. *Dentomaxillofacial Radiology*, 38 (5), 262–273.

Debelmas, A., Ketoff, S., Lanciaux, S., Corre, P., Friess, M., K, et al. (2019). Reproducibility assessment of Delaire cephalometric analysis using reconstructions from computed tomography. *Journal of Stomatology, Oral and Maxillofacial Surgery*, 121 (1), 35–39.

Dreyer, K. J., & Raymond Geis, J (2017). When machines think: Radiology's next frontier. *Radiology*, 285 (3), 713–718.

Durão, A. P. R., Morosolli, A., Pittayapat, P., Bolstad, N., Ferreira, A. P., et al. (2015). Cephalometric landmark variability among orthodontists and dentomaxillofacial radiologists: a comparative study. *Imaging Science in Dentistry*, 45 (4), 213–20.

Farooq, M. U., Khan, Mohd. A., Imran, S., Sameera, A., Qureshi, A., et al. (2016). Assessing the Reliability of Digitalized Cephalometric Analysis in Comparison with Manual Cephalometric Analysis. *Journal of Clinical and Diagnostic Research*, 10 (10), 20–23.

Forsting, M (2017). Machine Learning Will Change Medicine. *Journal of Nuclear Medicine*, 58 (3), 357–358.

Hung, K., Montalvao, C., Tanaka, R., Kawai, T., Bornstein, M. M (2019). The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiology*, 48 (20190107), 1-22, 2019.

Hwang, H.-W., Park, J.-H., Moon, J.-H., Yu, Y., Kim, H., H. et al. (2020). Automated identification of cephalometric landmarks: Part 2- Might it be better than human? *The Angle Orthodontist*, 90 (1), 69–76.

Khan, A., Javed, M. Q., Bilal, R., Gaikwad, R. N (2020). Retrospective quality assurance audit of Lateral Cephalometric Radiographs at postgraduate teaching hospital. *Pakistan Journal of Medical Sciences*, 36 (7), 1601-1606.

Khanagar, S. B., Al-Ehaideb, A., Maganur, P. C., Vishwanathaiah, S., Patil, S., et al. (2021). Developments, application, and performance of artificial intelligence in dentistry – A systematic review. *Journal of Dental Sciences*, 16 (1), 508–522.

Kunz, F., Stellzig-Eisenhauer, A., Zeman, F., Boldt, J (2019). Artificial intelligence in orthodontics: Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. *Journal of Orofacial Orthopedics / Fortschritte der Kieferorthopädie*, 81 (1), 52–68.

Landis, J. R., Koch, G. G (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–74.

Leonardi, R., Giordano, D., & Maiorana, F (2009). An evaluation of cellular neural networks for the automatic identification of cephalometric landmarks on digital images. *Journal of Biomedicine and Biotechnology*, (2009), 1-11.

Livas, C., Delli, K., Spijkervet, F. K. L., Vissink, A., Dijkstra, P. U (2019). Concurrent validity and reliability of cephalometric analysis using smartphone apps and computer software. *The Angle Orthodontist*, 89 (6), 889–896.

Mahto, R. K., Kafle, D., Giri, A., Luintel, S., Karki, A (2022). Evaluation of fully automated cephalometric measurements obtained from web-based artificial intelligence driven platform. *BMC Oral Health*, 22 (1), 1-8.

Masse, J.-F (2019). Will the orthodontic profession disappear? *Journal of Dental Sleep Medicine*, 6 (2), 1-2.

Mcnamara, A (1984). A method of cephalomettic evaluation. *American Journal of Orthodontics*, 6, 449-469.

Meriç, P., & Naoumova, J (2020). Web-based Fully Automated Cephalometric Analysis: Comparisons between App-aided, Computerized, and Manual Tracings. *Turkish Journal of Orthodontics*, 33 (3), 142–149.

Moon, J. H., Hwang, H. W., Yu, Y., Kim, M. G., Donatelli, R. E., L. …, S. J (2020). How much deep learning is enough for automatic identification to be reliable? A cephalometric example. *Angle Orthodontist*, 90 (6), 823–830.

Obermeyer, Z., & Emanuel, E. J (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine*, 375 (13), 1216–1219.

Olmez, H., Gorgulu, S., Akin, E., Bengi, A. O., Tekdemir, İ., Ors, F (2011). Measurement accuracy of a computer-assisted three-dimensional analysis and a conventional two-dimensional method. *The Angle Orthodontist*, 81 (3), 375–382.

Ongkosuwito, E. M., Katsaros, C., Van't Hof, M. A., Bodegom, J. C., Kuijpers-Jagtman, A. M (2002). The reproducibility of cephalometric measurements: a comparison of analogue and digital methods. *European Journal of Orthodontics*, 24, 655–665.

Park, J.-H., Hwang, H.-W., Moon, J.-H., Yu, Y., Kim, H., H. et al. (2019). Automated identification of cephalometric landmarks: Part 1—Comparisons between the latest deep-learning methods YOLOV3 and SSD. *The Angle Orthodontist*, 89 (6), 903–909.

Ravikumar, D., N., S., Ramakrishna, M., Sharna, N., Robindro, W (2019). Evaluation of McNamara's analysis in South Indian (Tamil Nadu) children between 8–12 years of age using lateral cephalograms. *Journal of Oral Biology and Craniofacial Research*, 9 (2), 193–197.

Santoro, M., Jarjoura, K., & Cangialosi, T. J (2006). Accuracy of digital and analogue cephalometric measurements assessed with the sandwich technique. *American Journal of Orthodontics and Dentofacial Orthopedics*, 129 (3), 345–351.

Shahidi, S., Oshagh, M., Gozin, F., Salehi, P., Danaei, S. M (2013). Accuracy of computerized automatic identification of cephalometric landmarks by a designed software. *Dentomaxillofacial Radiology*, 42, (1), p. 1-8.

Silva, T. P., Hughes, M. M., Menezes, L. Dos S., Melo, M. De F. B. De, Takeshita, W. M., Freitas, P. H. L. De (2021). Artificial Intelligence-Based Cephalometric Landmark Annotation and Measurements According to Arnett's Analysis: Can we trust a bot to do that? *Dentomaxillofacial Radiology*, 50, (20200548), 1-6.

Subramanian, A. K., Chen, Y., Almalki, A., Sivamurthy, G., Kafle, D (2022). Cephalometric Analysis in Orthodontics Using Artificial Intelligence—A Comprehensive Review. *BioMed Research International*, 2022, 1–9.

Yu, H. J., Cho, S. R., Kim, M. J., Kim, W. H., Kim, J. W., Choi, J (2020). Automated Skeletal Classification with Lateral Cephalometry Based on Artificial Intelligence. *Journal of Dental Research*, 99 (3), 249–256.

Zamrik, O. M., & Iseri, H (2021). The reliability and reproducibility of an Android cephalometric smartphone application in comparison with the conventional method. *Angle Orthodontist*, 91 (2), 236–242.