http://dx.doi.org/10.1590/0100-3984.2023.0056

Validação de algoritmo de aprendizado profundo para detecção da idade óssea em pacientes de São Paulo, Brasil

Validation of a deep learning algorithm for bone age estimation among patients in the city of São Paulo, Brazil

Augusto Sarquis Serpa[1,2,a], Abrahão Elias Neto[1,b], Felipe Campos Kitamura[1,2,c], Soraya Silveira Monteiro[1,d], Rodrigo Ragazzini[1,e], Gustavo Antunes Rodrigues Duarte[1,f], Lucas André Caricati[1,g], Nitamar Abdala[1,3,h]

1. Escola Paulista de Medicina da Universidade Federal de São Paulo (EPM-Unifesp), São Paulo, SP, Brasil. 2. Dasa, São Paulo, SP, Brasil. 3. Ionic Health, São José dos Campos, SP, Brasil.

Correspondência: Dr. Abrahão Elias Neto. Rua Apeninos, 236, ap. 111, Liberdade. São Paulo, SP, Brasil, 01533-000. E-mail: abrahao.elias@unifesp.br.

a. https://orcid.org/0000-0002-7292-9017; b. https://orcid.org/0009-0002-0655-7494; c. https://orcid.org/0000-0002-9992-5630; d. https://orcid.org/0009-0000-0081-4061; e. https://orcid.org/0000-0003-4200-2066; f. https://orcid.org/0009-0001-6556-8866; g. https://orcid.org/0000-0002-1149-083X; h. https://orcid.org/0000-0002-0421-0959.

Submetido em 29/5/2023. Revisado em 11/7/2023. Aceito em 31/7/2023.

Como citar este artigo:

Serpa AS, Elias Neto A, Kitamura FC, Monteiro SS, Ragazzini R, Duarte GAR, Caricati LA, Abdala N. Validação de algoritmo de aprendizado profundo para detecção da idade óssea em pacientes de São Paulo, Brasil. Radiol Bras. 2023 Set/Out;56(5):263–268.

Resumo **Objetivo:** Validar em indivíduos paulistas um modelo de aprendizado profundo (*deep learning* – DL) para estimativa da idade óssea, comparando-o com o método de Greulich e Pyle.

Materiais e Métodos: Estudo transversal com radiografias de mão e punho para idade óssea. A análise manual foi feita por um radiologista experiente. Foi usado um modelo baseado em uma rede neural convolucional que ficou em terceiro lugar no desafio de 2017 da Radiological Society of North America. Calcularam-se o erro médio absoluto (*mean absolute error* – MAE) e a raiz do erro médio quadrado (*root mean-square error* – RMSE) do modelo contra o radiologista, com comparações entre sexo, etnia e idade.

Resultados: A amostra compreendia 714 exames. Houve correlação entre ambos os métodos com coeficiente de determinação de 0,94. O MAE das predições foi 7,68 meses e a RMSE foi 10,27 meses. Não houve diferenças estatisticamente significantes entre sexos ou raças ($p > 0,05$). O algoritmo superestimou a idade óssea nos mais jovens ($p = 0,001$).

Conclusão: O nosso algoritmo de DL demonstrou potencial para estimar a idade óssea em indivíduos paulistas, independentemente do sexo e da raça. Entretanto, há necessidade de aprimoramentos, particularmente em pacientes mais jovens.

Unitermos: Inteligência artificial; Aprendizado de máquina; Aprendizado profundo; Desenvolvimento ósseo; Crescimento.

Abstract **Objective:** To validate a deep learning (DL) model for bone age estimation in individuals in the city of São Paulo, comparing it with the Greulich and Pyle method.

Materials and Methods: This was a cross-sectional study of hand and wrist radiographs obtained for the determination of bone age. The manual analysis was performed by an experienced radiologist. The model used was based on a convolutional neural network that placed third in the 2017 Radiological Society of North America challenge. The mean absolute error (MAE) and the root-mean-square error (RMSE) were calculated for the model versus the radiologist, with comparisons by sex, race, and age.

Results: The sample comprised 714 examinations. There was a correlation between the two methods, with a coefficient of determination of 0.94. The MAE of the predictions was 7.68 months, and the RMSE was 10.27 months. There were no statistically significant differences between sexes or among races ($p > 0.05$). The algorithm overestimated bone age in younger individuals ($p = 0.001$).

Conclusion: Our DL algorithm demonstrated potential for estimating bone age in individuals in the city of São Paulo, regardless of sex and race. However, improvements are needed, particularly in relation to its use in younger patients.

Keywords: Artificial intelligence; Machine learning; Deep learning; Bone development; Growth.

INTRODUÇÃO

A determinação precisa da idade óssea desempenha papel vital no monitoramento do desenvolvimento ósseo, agindo como um indicador fiel da idade biológica e do prognóstico de crescimento[1]. Existem diversos métodos manuais de estimativa de idade óssea que utilizam radiografias de várias partes do corpo[2]. Contudo, a mão e o punho são os mais frequentemente escolhidos, graças à presença de múltiplos centros de ossificação, simplicidade da técnica, proteção radiológica adequada e baixo custo do procedimento[1]. Preconiza-se a utilização do membro esquerdo por algumas razões, dentre elas o fato de que a maioria das pessoas é destra, portanto, há maior chance de lesões ósseas neste lado[1].

Entre os métodos radiográficos empregados para a avaliação da idade óssea, o de Greulich e Pyle[3] é o mais

utilizado[2] e envolve a análise dos centros de ossificação da mão e do punho esquerdo em comparação com um atlas de imagens padrão. No entanto, esse método foi originalmente desenvolvido com base em uma população caucasiana e norte-americana na década de 1950[3]. Portanto, sua aplicabilidade e precisão, quando utilizado em outras populações, têm sido questionadas[4]. Além disso, há controvérsia na literatura quanto à sua reprodutibilidade, com discrepâncias significativas entre os resultados dos estudos que se propuseram a avaliar a variabilidade intraobservador e interobservador das leituras[5].

Diante desse cenário, foram propostos diversos modelos automatizados que utilizam a inteligência artificial (IA) para detecção da idade óssea, a maioria deles baseados em aprendizado de máquina (*machine learning* – ML) tradicional, sendo o BoneXpert o mais amplamente empregado[6,7]. No entanto, a maioria desses algoritmos foi desenvolvida com base em populações oriundas dos Estados Unidos e da Europa Ocidental, e poucos estudos consideraram as particularidades étnicas e socioeconômicas dos indivíduos na análise dos resultados[6].

Até o momento, não há estudos que avaliem o desempenho desses algoritmos na população brasileira. Diante dessa lacuna, este estudo tem como objetivo validar, em crianças e adolescentes de São Paulo, as predições de um modelo baseado em aprendizado profundo (*deep learning* – DL), um subtipo de ML, para a estimativa da idade óssea, comparando os resultados obtidos com a análise realizada por um radiologista treinado, por meio do método Greulich e Pyle. Esta validação local é essencial para garantir a precisão e relevância clínica desses modelos de IA antes de sua implementação em larga escala no cenário clínico brasileiro.

MATERIAIS E MÉTODOS

Foi realizado um estudo transversal com radiografias de mão e punho esquerdos do nosso serviço, registradas como exames para idade óssea. O estudo foi aprovado pelo comitê de ética em pesquisa com base no projeto de pesquisa "Desenvolvimento de bancos de dados de imagens médicas para fomentar pesquisas e desafios de *machine learning* na radiologia".

Foi criado um banco de dados com as radiografias realizadas entre 2018 e 2022. Os critérios de inclusão foram pacientes de ambos os sexos e a presença de laudo descrevendo a idade óssea. Todos os exames foram laudados por um radiologista com experiência de três anos em idade óssea, pelo método Greulich e Pyle. Foram excluídos exames bilaterais, de outras partes do corpo que estavam registrados incorretamente, os com técnica e/ou posicionamento inadequados, os com cateteres periféricos e os com deformidades ósseas que prejudicavam a análise. Foi utilizada uma amostra de conveniência pelo fato de não existir um método de cálculo amostral universalmente aceito para modelos de DL. Os exames foram tornados anônimos por meio de *software* específico (RSNA Anonymizer), cujo *download* e código fonte acham-se disponíveis em http://mirc.rsna.org/download/Anonymizer-installer.jar. Em seguida, as imagens foram enviadas para um *software* de anotação de imagens médicas baseado na nuvem (MD.ai – New York, NY). Por intermédio deste, foram anotadas quais radiografias possuíam um ou mais critérios de exclusão, para posterior eliminação do estudo. Também foram anotados dados sobre idade, sexo, cor, idade cronológica e idade óssea laudada de cada paciente, mediante análise de prontuário. Foram consideradas as cinco raças propostas pela classificação do Instituto Brasileiro de Geografia e Estatística, de acordo com o último censo publicado: amarela, branca, indígena, parda e preta[8].

Após a coleta e anotação dos dados, foi realizada a inferência dos exames em um modelo de DL baseado em uma rede neural convolucional (Figura 1), desenvolvido pela Universidade Federal de São Paulo em parceria com a Universidade Federal de Goiás. Na etapa de treinamento desse algoritmo, foi realizada divisão da base de dados em cinco grupos e feita validação cruzada. A predição final envolve a média aritmética dos quatro modelos que tiveram o melhor resultado individual. Os hiperparâmetros foram os seguintes: taxa de aprendizado inicial de 10-4, tamanho do *batch* de 16 e número de épocas de 100. Foi usado o otimizador Adam. Como pré-processamento das imagens, todos os *pixels* são divididos por 255, de modo que fiquem no intervalo [0, 1], e em seguida normalizados utilizando-se a média e desvio-padrão de cada exame. Em seguida, há um redimensionamento da imagem para o tamanho 550 × 550, preservando-se as proporções originais, e se necessário, é realizado *padding* com zeros às bordas da imagem. Tais etapas de pré-processamento também foram utilizadas em todas as radiografias incluídas no presente estudo. Na fase de treinamento, também foi realizada ampliação de dados em uma porcentagem dos exames, com modificações como rotação de ± 30°, inversão no eixo horizontal e *zoom* de ± 10%. Esse modelo foi previamente treinado e testado na base de dados da competição de IA da Radiological Society of North America (RSNA) em 2017, tendo alcançado o terceiro lugar entre equipes participantes de todo o mundo, com erro médio absoluto (*mean absolute error* – MAE) de 4,38 meses[9].

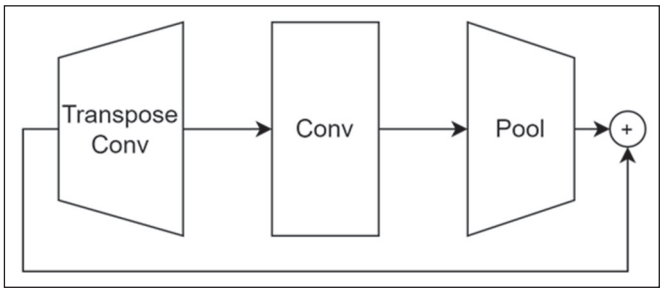

Figura 1. Arquitetura *ice module*, bloco básico do modelo utilizado, que consiste em uma convolução transposta seguida por uma camada de convolução e *pooling*, além de um atalho por meio de uma conexão residual.

A etapa seguinte foi uma análise comparativa com o laudo do radiologista. Foram calculados os erros absolutos do resultado do algoritmo em relação à leitura convencional para cada paciente, e o resultado expressado em meses. Com isso, foi calculado o MAE, cujo cálculo consiste na soma dos erros absolutos, dividida pelo número total de exames. Essa métrica é bastante utilizada na avaliação do desempenho de algoritmos de IA que envolvem predições de variáveis numéricas. A sua vantagem em relação a outras métricas similares, como a raiz do erro médio quadrado (*root mean-square error* – RMSE), é que não está sujeita a variações na distribuição da magnitude dos erros e no tamanho da amostra[10]. No entanto, pelo fato de outros estudos similares usarem a RMSE, também calculamos essa métrica para toda a amostra, com o intuito de permitir análises comparativas com tais estudos.

Em razão de as variáveis de interesse não possuírem distribuição normal, as análises descritivas também foram realizadas por mediana e intervalo interquartil (IIQ). Para detecção dos pontos fora da curva, utilizou-se o cálculo da mediana + 1,5 × IIQ. Foram feitas análises estatísticas comparativas dos resultados por sexo e por grupo etário pelo teste de Mann-Whitney. Já para comparação racial foi aplicado o teste de Kruskal-Wallis em todos os grupos. Foi também realizada uma regressão linear entre a idade óssea laudada e a predição do algoritmo, com cálculo do coeficiente de correlação de Pearson e coeficiente de determinação. Foi traçado o gráfico de Bland-Altman para estudo do erro não absoluto, que preserva a informação se o modelo superestimou ou subestimou a idade óssea em relação ao radiologista.

O estudo foi realizado com a linguagem de programação Python versão 3[11], sendo utilizadas as bibliotecas Pandas[12] e SciPy[13] para as análises estatísticas. Já para a elaboração dos gráficos, foram usadas as bibliotecas Mat PlotLib[14] e Seaborn[15]. Para a inferência do algoritmo, foram utilizados os pacotes PyTorch[16] e NumPy[17]. Em todas as conclusões obtidas pelas análises inferenciais foi utilizado o nível de significância de 5% ($p \leq 0,05$).

RESULTADOS

No total, foram elegíveis 764 exames que preenchiam os critérios de inclusão. Destes, 50 foram eliminados por preencherem algum dos critérios de exclusão, restando 714 exames (Figura 2). Os dados demográficos dos 714 exames incluídos estão apresentados na Tabela 1. As idades cronológicas mínima e máxima foram 1 ano e 3 meses e 19 anos e 10 meses, respectivamente, e apenas seis

Tabela 1—Características gerais da amostra estudada.

Variável	(N = 714)
Sexo, n (%)	
Feminino	369 (51,68)
Masculino	345 (48,32)
Idade cronológica (anos), mediana (IIQ)	10,79 (8,27–13,33)
Idade óssea (anos), mediana (IIQ)	11 (8,83–13,5)
Cor, n (%)	
Branca	338 (47,34)
Parda	214 (29,97)
Preta	23 (3,22)
Amarela	1 (0,14)
Indígena	1 (0,14)
Sem informação	137 (19,19)

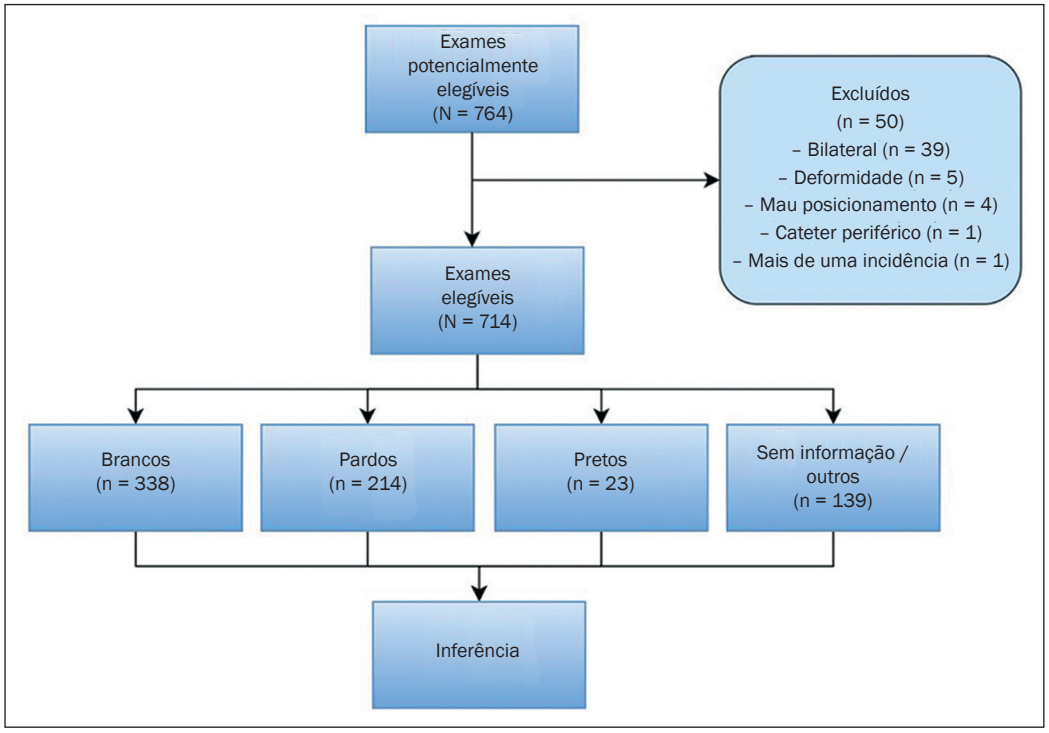

Figura 2. Fluxograma do estudo.

pacientes tinham idade inferior a 3 anos completos. De 137 pacientes não havia informações sobre cor, os quais foram excluídos das análises referentes a essa variável independente. Havia apenas um paciente da cor amarela e um paciente indígena, que também foram excluídos da análise pela quantidade insuficiente de casos.

No estudo da correlação entre a idade óssea laudada pelo radiologista e a predição do algoritmo, obteve-se uma regressão linear cuja reta aproximou-se da reta ideal, que seria o caso em que o modelo acertasse todas as predições (y = x). O coeficiente de correlação de Pearson foi 0,97 e o coeficiente de determinação, 0,94 (Figura 3). A análise da correlação linear e do gráfico de Bland-Altman (Figura 4), que estuda o erro não absoluto (predição – idade óssea laudada), sugere uma tendência de o algoritmo superestimar a idade óssea nos mais novos.

O MAE das predições em relação à idade óssea laudada foi 7,68 meses para toda a amostra. Já a RMSE foi 10,27 meses (0,86 anos). A Tabela 2 descreve os MAEs, medianas e IIQs para todos os exames e discriminados por sexo, por cor e por grupo etário e os respectivos valores de *p* das análises inferenciais entre os grupos. A divisão do grupo etário foi feita no percentil 50 para idade cronológica, para se confirmar a hipótese de que o modelo superestimou a idade óssea nos mais jovens, o que foi comprovado. Já a comparação entre os sexos e entre as raças não revelou diferença estatisticamente significante (Figuras 5 e 6). Mediante análise do IIQ, foram encontrados 19 pontos fora da curva, explicitados nos gráficos de caixa como as representações em formato de losango (Figuras 5 e 6).

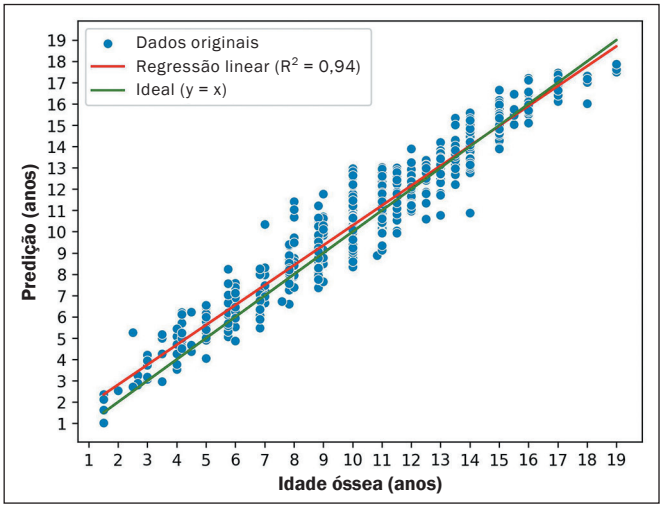

Figura 3. Gráfico das predições em relação à idade óssea laudada.

DISCUSSÃO

Neste estudo buscamos validar um algoritmo de DL para cálculo da idade óssea a partir de radiografias de mãos e punhos de pacientes acompanhados no nosso serviço. O MAE da predição do modelo em relação ao laudo do radiologista foi 7,68 meses, valor inferior ao de 9,96 meses encontrado em uma meta-análise recente entre estudos que utilizaram técnicas diversas de ML para predição de idade óssea[6]. Isso indica que o desempenho do algoritmo é consistente ou melhor que o de outros modelos propostos em um contexto clínico real. No entanto, o desempenho nos nossos dados foi consideravelmente pior

Tabela 2—Análise do MAE, por sexo, cor e idade, em meses.

Variável	MAE	Mediana	IIQ	*P*
Sexo (N = 714)				0,575
Feminino	7,55	5,88	3,05–10,65	
Masculino	7,82	5,64	2,36–11,34	
Cor (n = 575)				0,368†
Branca	7,25	5,50	2,21–10,36	
Parda	7,85	6,13	2,86–11,61	
Preta	6,32	6,52	4,22–8,46	
Idade* (N = 714)				0,001‡
≤ 10,79 anos	8,43	6,42	3,23–11,87	
> 10,79 anos	6,41	5,16	2,19–9,62	

* Percentil 50 para idade cronológica. † Teste de Kruskal-Wallis simultâneo para os três grupos. ‡ Estatisticamente significante.

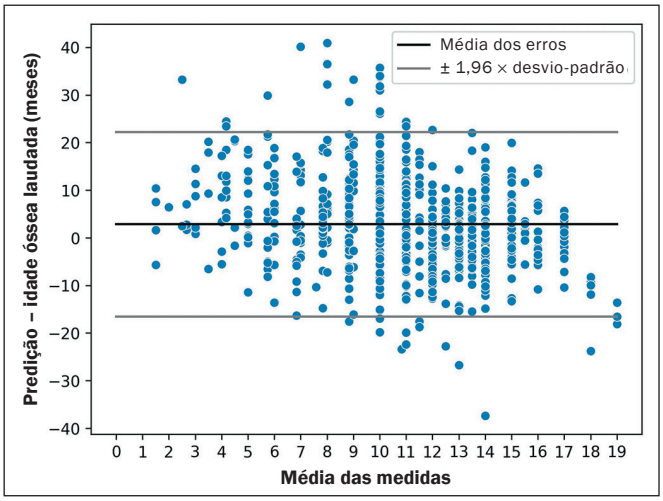

Figura 4. Gráfico de Bland-Altman com o erro em relação à média das medidas manuais e do algoritmo.

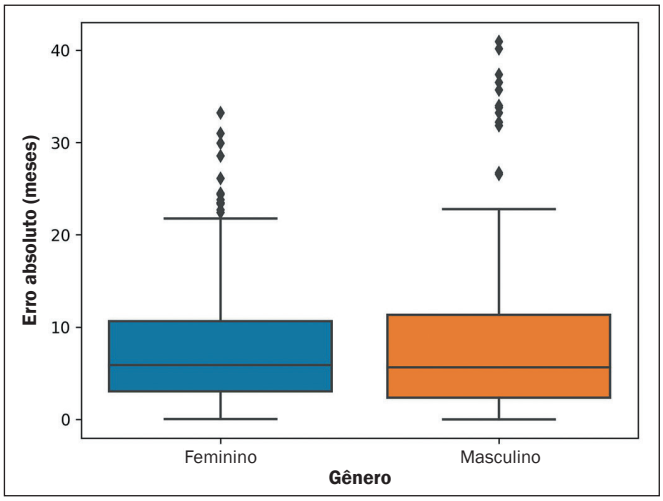

Figura 5. Diagramas de caixas dos erros absolutos em relação a cada sexo da amostra estudada.

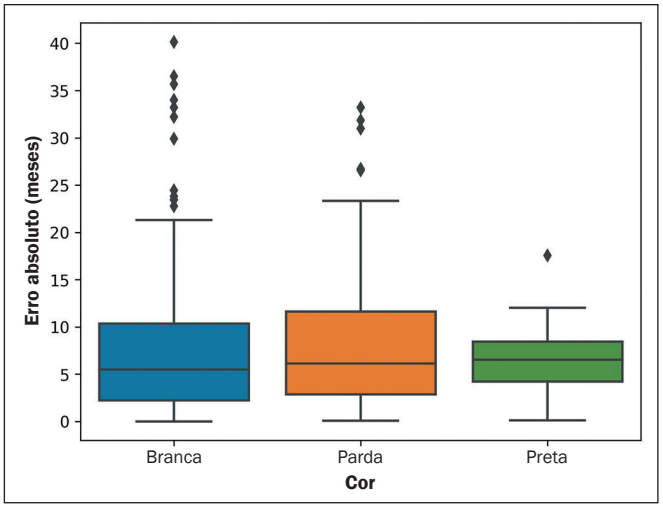

Figura 6. Diagramas de caixas dos erros absolutos em relação a cada cor da amostra estudada.

em relação ao desafio da RSNA (MAE de 4,38 meses)[9]. Uma possível explicação é o fato de o algoritmo ter sido treinado em populações oriundas da América do Norte, com fenótipos diferentes da população brasileira. Todavia, vale ressaltar que o grupo de teste, no banco de dados do desafio, foi anotado com base na opinião de seis radiologistas, o que reduz o erro humano e pode explicar parte dessa discrepância entre os MAEs.

A RMSE foi 10,27 meses (0,86 anos) para toda a amostra. O algoritmo mais amplamente empregado e validado, o BoneXpert, baseado em ML tradicional, obteve um RMSE de 0,72 anos na sua versão 2[18–20] e de 0,62 anos na sua versão 3[7], ambas aferidas em um grupo de teste, independente do grupo de treino, de pacientes de Tübingen, Alemanha, e baseado na leitura de apenas um radiologista. Apesar do nosso pior desempenho, deve-se ressaltar que o treinamento desse outro modelo foi em pacientes de origem europeia ou norte-americana, cujas características étnicas, socioeconômicas e nutricionais aproximam-se mais do grupo de teste utilizado do que no nosso estudo. Ao compararmos com um modelo que usou DL, também treinado no banco de dados da RSNA e validado em um banco de dados externo, verificamos um desempenho pior do nosso algoritmo[21] (EMA de 5,96 meses). Porém, essa validação foi realizada apenas em centros dos Estados Unidos, mesmo país de origem dos exames em que foi treinado, e a anotação foi feita por quatro radiologistas, reduzindo substancialmente o erro humano.

A ausência de diferença estatisticamente significante entre os sexos e entre as raças em relação ao erro absoluto sugere que o modelo tem um desempenho uniforme para meninos e meninas de diferentes etnias, o que é uma característica desejável para aplicação clínica. Atualmente, poucos estudos na literatura comparam o desempenho de algoritmos de idade óssea entre diferentes raças[6]. Apesar disso, é importante salientar que o baixo número de indivíduos pretos na amostra pode ter sido insuficiente para o teste estatístico capturar alguma diferença e não refletiu a distribuição étnica da população brasileira[22].

O nosso método automático superestimou a idade óssea nos pacientes mais jovens. Uma possível explicação para esse achado é a maior variabilidade no desenvolvimento ósseo nesses indivíduos, o que pode ser difícil para o modelo capturar[23]. Nesse sentido, aprimoramentos no treinamento do modelo podem ser necessários para melhorar o desempenho do algoritmo nessa faixa etária.

Apesar dos resultados encorajadores, nosso estudo apresenta várias limitações que devem ser consideradas ao interpretar os achados. Novamente, a idade óssea foi determinada com base na avaliação de apenas um radiologista, o que introduz um erro esperado, já que a interpretação dos métodos tradicionais de idade óssea é subjetiva e pode variar entre diferentes observadores[5,23]. Além disso, a nossa amostra contou com um baixo número de participantes autodeclarados como pretos, e apenas um indivíduo de cada categoria autodeclarada como amarelo e indígena, bem como seis pacientes com idade abaixo de três anos, o que pode ter impactado a generalização dos nossos resultados. A falta de informações sobre a raça de alguns pacientes também é uma lacuna significativa. Por fim, a ausência de informações detalhadas sobre as comorbidades de cada paciente é outra limitação, uma vez que certas condições médicas podem influenciar o desenvolvimento ósseo[23].

Em conclusão, o algoritmo de DL validado no presente estudo apresenta um desempenho promissor para estimar a idade óssea em crianças e adolescentes na população brasileira de ambos os sexos e entre diferentes raças. No entanto, é importante considerar suas limitações e a necessidade de refinamento para melhorar sua aplicabilidade clínica, especialmente em pacientes mais jovens. Além disso, o algoritmo não deve ser visto como um substituto para a avaliação do radiologista, mas sim como uma ferramenta complementar no processo de determinação da idade óssea.

Agradecimento

A Ernandez Rodrigues dos Santos, nosso administrador do PACS, por ajudar na coleta de dados.

REFERÊNCIAS

1. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol. 2015;24:143–52.
2. Breen MA, Tsai A, Stamm A, et al. Bone age assessment practices in infants and older children among Society for Pediatric Radiology members. Pediatr Radiol. 2016;46:1269–74.
3. Bayer LM. Radiographic atlas of skeletal development of the hand and wrist. Calif Med. 1959;91:53.
4. Alshamrani K, Messina F, Offiah AC. Is the Greulich and Pyle atlas applicable to all ethnicities? A systematic review and meta-analysis. Eur Radiol. 2019;29:2910–23.
5. Berst MJ, Dolan L, Bogdanowicz MM, et al. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards. AJR Am J Roentgenol. 2001; 176:507–10.
6. Dallora AL, Anderberg P, Kvist O, et al. Bone age assessment with various machine learning techniques: a systematic literature review and meta-analysis. PLoS One. 2019;14:e0220242.

7. Martin DD, Calder AD, Ranke MB, et al. Accuracy and self-validation of automated bone age determination. Sci Rep. 2022;12:6388.
8. Instituto Brasileiro de Geografia e Estatística. Censo brasileiro de 2010. Rio de Janeiro, RJ: IBGE; 2012.
9. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. Radiology. 2019; 290:498–503.
10. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005;30:79–82.
11. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley, CA: CreateSpace; 2009.
12. The pandas development team. Pandas-dev/pandas: Pandas (v2.0.1). Zenodo; 2023.
13. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17:261–72.
14. Hunter JD. Matplotlib: a 2D graphics environment. Computing in Science & Engineering. 2007;9:90–5.
15. Waskom ML. Seaborn: statistical data visualization. J Open Source Soft. 2021;6:3021.
16. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. arXiv:1912.01703v1.
17. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585:357–62.
18. Martin DD, Deusch D, Schweizer R, et al. Clinical application of automated Greulich-Pyle bone age determination in children with short stature. Pediatr Radiol. 2009;39:598–607.
19. Martin DD, Heil K, Heckmann C, et al. Validation of automatic bone age determination in children with congenital adrenal hyperplasia. Pediatr Radiol. 2013;43:1615–21.
20. Martin DD, Meister K, Schweizer R, et al. Validation of automatic bone age rating in children with precocious and early puberty. J Pediatr Endocrinol Metab. 2011;24:1009–14.
21. Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. Radiology. 2021:301:692–9.
22. Instituto Brasileiro de Geografia e Estatística. Pesquisa nacional por amostra de domicílios contínua. Rio de Janeiro, RJ: IBGE; 2021.
23. Cavallo F, Mohn A, Chiarelli F, et al. Evaluation of bone age in children: a mini-review. Front Pediatr. 2021;9:580314.