

MARCELO FROEDER ARCURI

**UM *CHAT BOT* ASSOCIADO A UM ALGORITMO DE APRENDIZADO
DE MÁQUINA PARA PREDIÇÃO DE IDADE ÓSSEA PEDIÁTRICA
ATRAVÉS DE RADIOGRAFIAS DE MÃO E PUNHO**

Orientador: Prof. Dr. Nitamar Abdala

Coorientador: Prof. Dr. Felipe Kitamura

Tese apresentada à Universidade Federal
de São Paulo – Escola Paulista de Medicina
para obtenção do Título de Mestre/ Doutor
em Ciências.

SÃO PAULO

2020

Arcuri, Marcelo Froeder

Um *chat bot* associado a um algoritmo de aprendizado de máquina para predição de idade óssea pediátrica através de radiografias de mão e punho /
Marcelo Froeder Arcuri. -- São Paulo, 2020.

Dissertação (Trabalho de Conclusão de Curso) – Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Residência Médica em Radiologia e Diagnóstico por Imagem.

Título em inglês: A chat bot with a machine learning algorithm for predicting pediatric bone age through hand and wrist radiographs.

1. Idade óssea. 2. Desenvolvimento ósseo. 3. Transtornos do crescimento ósseo. 4. Deep Learning. 5. Inteligência Artificial.

**UNIVERSIDADE FEDERAL DE SÃO PAULO
ESCOLA PAULISTA DE MEDICINA**

DEPARTAMENTO DE DIAGNÓSTICO POR IMAGEM

Chefe do Departamento:

Prof. Dr. Henrique Carrete Junior

Supervisor da Residência Médica:

Prof. Dr. Rodrigo Regacini

AGRADECIMENTOS

Ao Dr. Marcelo Straus Takahashi, à Dra. Taisa Gasparetto e a toda a equipe de inteligência artificial e inovação do Dasa pela colaboração.

SUMÁRIO

1. RESUMO	1
2. INTRODUÇÃO	3
3. MATERIAIS E MÉTODOS	5
4. RESULTADOS	7
5. DISCUSSÃO	9
6. CONCLUSÃO	10
7. REFERÊNCIAS	11
8. ANEXOS	12

1. RESUMO

Objetivo: avaliar se a análise da idade óssea feita por um algoritmo de inteligência artificial corresponde à análise dos médicos radiologistas experientes. **Métodos:** Foram selecionados exames consecutivos de pacientes pediátricos realizados entre Janeiro e Agosto de 2019 laudados por dois radiologistas especializados em pediatria, sendo excluídos exames com erro técnico ou de registro, esses exames foram divididos em dois grupos de acordo com o sexo do paciente. As interpretações dos radiologistas foram comparadas entre si. Foi feita a inferência da idade óssea através de um algoritmo de *deep learning* com entrada dos arquivos DICOM e comparada com a média do resultado dos médicos. Também foi desenvolvido uma ferramenta de chat bot disponível online através do qual foram usadas fotografias digitais (das mesmas radiografias usadas nos outros grupos) como entrada ao algoritmo, sendo o desempenho desta interpretação comparada com os resultados dos radiologistas e com a interpretação dos arquivos DICOM. **Resultados:** Não houve diferença estatisticamente significativa entre os dois radiologistas ou entre a média destes e o grupo com entrada por DICOM para o sexo feminino. No entanto, houve diferença estatisticamente significativa entre o grupo com entrada por fotografia e a média dos radiologistas, assim como entre esse grupo e o grupo DICOM. Surpreendentemente houve diferença estatisticamente significativa entre a média dos radiologistas e o grupo com entrada por DICOM para o sexo masculino. **Conclusão:** O desenvolvimento de ferramentas como esta demonstram o potencial de algoritmos de IA aplicados à radiologia e diagnóstico por imagem, de uma maneira prática, compreensível e de fácil acesso, ilustrando futuras aplicações desta tecnologia. O processo de validação clínica é essencial para identificarmos e corrigirmos situações inesperadas como o maior erro do algoritmo para o sexo masculino.

ABSTRACT

Purpose: to compare something and anything. **Methods:** the experimental study consisted in something and anything. **Results:** the results of something and anything show statistical difference between the two things. **Conclusion:** something show statistical difference in relation to anything.

2. INTRODUÇÃO

A mensuração da maturidade esquelética, ou idade óssea, é importante para a análise do crescimento e dos distúrbios de crescimento na população pediátrica. Vários métodos estão disponíveis para a avaliação da idade óssea. Todos são baseados no reconhecimento de indicadores de maturidade; mudanças progressivas no aspecto radiográfico dos ossos da mão e punho, com desenvolvimento dos centros de ossificação. O método mais usado no mundo é o Atlas de Greulich e Pyle, no qual o exame do paciente é manualmente comparado com padrões de normalidade. (1)

No entanto, a interpretação manual destes exames apresenta uma elevada taxa de variabilidade inter e intraobservador, além de ser considerada pouco estimulante e moroso por radiologistas, sendo um dos primeiros exames a ter sido candidato para automação (2).

Com o advento das novas técnicas de inteligência artificial (notadamente o *deep learning* e as redes neurais convolucionais), o surgimento de grandes bancos de dados e unidades de processamento gráfico mais poderosas nos últimos anos, houve um aumento significativo na capacidade de algoritmos reconhecerem imagens com alta acurácia (inclusive em alguns casos melhor que seres humanos, desde 2015).

Estas técnicas tem ampla aplicação à radiologia e diagnóstico por imagem e introduziram uma outra área do conhecimento à realidade do médico radiologista. (3 e 4)

A avaliação automática de radiografias de mão e punho para determinar a idade óssea, se tornou uma aplicação precoce de *deep learning* na radiologia por preencher critérios que facilitaram este processo, como usar uma única imagem bidimensional (à semelhança dos modelos dos principais algoritmos desenvolvidos para interpretação e classificação de imagens), com referência padronizada (atlas), e um resultado único quantitativo (estimativa da idade óssea), que tem sido usada por várias décadas (disponibilizando grande quantidade de dados) (4).

Vários trabalhos científicos foram publicados sobre a investigação destes métodos e a performance dos modelos de *deep learning* em determinar a idade óssea, demonstrando ser comparável à de radiologistas experientes e aos modelos automáticos existentes (4 e 5). Inclusive, este tipo de exame foi o tema do primeiro desafio de inteligência artificial, no RSNA de 2017. (6).

Este fato incentivou a criação de muitos *softwares* para aplicação destes algoritmos. Neste cenário, o desenvolvimento de ferramentas práticas e facilmente acessíveis que empregam esta tecnologia poderia trazer um grande benefício para pacientes, serviços de imagem e médicos solicitantes, aumentando a acurácia e consistência do diagnóstico, reduzindo o tempo para sua realização e aumentando a disponibilidade dos radiologistas para tarefas mais exigentes.

3. MATERIAIS E MÉTODOS

Foram selecionados exames de radiografia de mão e punho para determinação de idade óssea de pacientes pediátricos realizados entre Janeiro e Agosto de 2019 nos serviços da Dasa. Aqueles que apresentaram erro técnico que pudesse alterar ou impedir a interpretação (por exemplo artefatos de movimento ou não inclusão de regiões anatômicas relevantes na radiografia) foram excluídos. Também foi excluído um exame com erro de registro, no qual o sexo do paciente estava trocado. Estes exames foram divididos em dois grupos de acordo com o sexo, sendo o grupo masculino composto por 353 radiografias e o feminino por 327, formando um total de 680 exames.

Todos os exames foram independentemente laudados por dois radiologistas especializados em pediatria usando como método a comparação com as imagens do atlas de Greulich Pyle.

As interpretações dos radiologistas foram comparadas entre si através de um test T de student pareado com duas caudas.

Os arquivos DICOM dos exames de radiografia de mão e punho foram anonimizados e usados como entrada em um algoritmo de inteligência artificial previamente treinado para esta finalidade.

Este algoritmo é composto por uma rede neural convolucional com múltiplas camadas com arquitetura no Modelo Ice modificado, sendo composto por uma camada de entrada com tamanho 550 x 550 pixels, uma convolução de transposição seguido por nove camadas de convolução e pooling ultrapassadas por uma conexão residual. Esta parte convolucional é seguida por duas camadas densas que tem como entrada o sexo do paciente e as características da imagem extraídos pela parte convolucional. A camada da saída é um neurônio com função de ativação ReLU, por se tratar de um modelo de regressão com saída de números positivos. O resultado é portanto um valor numérico representando a idade óssea em meses.

No treinamento foi utilizado *data augmentation* com rotação, zoom, deslocamento XY, brilho e contraste. Foram treinados 4 modelos, cada um com um conjunto de validação cruzada, utilizando otimizador Adam, com taxa de aprendizado de 0.0001. A função de custo utilizada foi o Mean Absolute Difference (MAD).

Os resultados foram convertidos ao valor mais próximo encontrado no atlas de Greulich Pyle para padronização dos dados e foram comparados com a média dos radiologistas através de um test T de student pareado com duas caudas.

Foram feitas fotografias digitais dos exames de radiografia de mão e punho usando dois tipos de aparelho celular: um Samsung Galaxy S9 e um Iphone 8, ambos com câmera de 12 megapixels. A única técnica definida para realizar a fotografia era incluir todo o exame no enquadramento, essa liberdade na aquisição da foto pretendia simular o uso real.

Estas fotografias foram usadas como entrada do algoritmo através de um *chatbot*: um software para realizar conversas em texto e mídia que simula uma interação humana. Este programa foi desenvolvido através da interface de programação de aplicações do Telegram (um serviço de mensagem instantânea) baseada em HTTP e tem como plataforma o serviço de nuvem da Amazon Web Service (AWS). O *bot* lê a entrada do usuário através do aplicativo do Telegram Messenger, se esta for uma imagem, é classificada através da inferência do modelo de inteligência artificial acima descrito. A saída é enviada através do Telegram e exibida como idade óssea através de texto.

Um mapa de ativação de classe ponderado em gradientes (Grad-CAM) também é gerado e plotado sobre a imagem original, identificando as regiões na radiografia importantes para a predição através de cores quentes, sendo exibida como uma imagem através do Telegram.

Também foi desenvolvido uma ferramenta de chat bot disponível online através do qual foram usadas fotografias digitais (das mesmas radiografias usadas nos outros grupos) como entrada ao algoritmo, sendo o desempenho desta interpretação comparada com os resultados dos radiologistas e com a interpretação dos arquivos DICOM pelo algoritmo, todos usando um teste T de student não pareado de duas caudas.

4. RESULTADOS

Para a população feminina não houve diferença estatisticamente significativa entre os dois radiologistas ou entre o grupo com DICOM como entrada no algoritmo e a média dos radiologistas.

No entanto, houve diferença estatisticamente significativa entre o grupo com fotografia digital como entrada no algoritmo e a média dos radiologistas. Também houve diferença estatisticamente significativa entre esse grupo (fotografia) e aquele com DICOM como entrada no algoritmo.

Tabela 1 – RESULTADOS PARA A POPULAÇÃO FEMININA

Grupos comparados (teste t de Student)	Valor de p
Radiologista1 vs Radiologista2	0,831261138
Média (Rad1 e Rad2) vs DICOM	0,975176227
Média (Rad1 e Rad2) vs fotografia	< 0,0001
fotografia vs DICOM	< 0,0001

Na população masculina novamente não houve diferença estatisticamente significativa entre os dois radiologistas.

Também houve diferença estatisticamente significativa quando o grupo que tinha fotografia como entrada foi comparada com a média dos radiologistas, assim como com o grupo com DICOM como entrada.

Tabela 2 – RESULTADOS PARA A POPULAÇÃO MASCULINA

Grupos comparados	Valor de p
Radiologista1 vs Radiologista2	0,195455487
Média (Rad1 e Rad2) vs DICOM	< 0,0001
Média (Rad1 e Rad2) vs Fotografia	< 0,0001
Fotografia vs DICOM	0,004481743

Supreendentemente houve diferença estatisticamente significativa entre o grupo com DICOM como entrada no algoritmo e a média dos radiologistas.

5. DISCUSSÃO

Era esperado que o resultado dos radiologistas demonstrasse certa equivalência estatística entre si, o que também era esperado entre os resultados da análise das imagens a partir dos arquivos DICOM e os resultados dos radiologistas. No entanto esta equivalência do algoritmo com os humanos se confirmou apenas para a população feminina e não para a masculina, o que exigiu uma explicação. Nos estimamos que as ponderações dos *perceptrons* da rede neural estavam calibradas com os dados da população feminina, sem correção para a população masculina, o que justificaria os resultados. Assim, modificações no código do algoritmo, novos testes e validações (possivelmente até novos treinos) seriam necessários para determinar a real explicação.

O uso de fotografias digitais adquiridas através de telefone celular como entrada para este modelo em ambos os sexos se provou inadequado. Isto implica que a qualidade da imagem de entrada interfere diretamente com a habilidade do algoritmo em determinar uma idade óssea com boa acurácia.

Soluções para este problema podem ser avaliadas através de mais estudos: com algoritmos mais refinados, capazes de interpretar imagens de menor qualidade (fotografias de celular) ou melhorar a qualidade da imagem de entrada, por exemplo através de imagens digitais recortadas.

6. CONCLUSÃO

Este estudo demonstrou que um algoritmo de inteligência artificial não apresentou diferença estatisticamente significativa com radiologistas experientes na determinação da idade óssea baseado em radiografias de mão e punho para a população feminina.

Apesar do resultado inesperado para a população masculina, evidencia-se o potencial de algoritmos de inteligência artificial aplicados à radiologia e diagnóstico por imagem, notadamente através do desenvolvimento de ferramentas como o *chat bot*, que demonstram de uma maneira prática, compreensível e de fácil acesso, ilustrando futuras aplicações desta tecnologia.

O processo de validação clínica é essencial para identificarmos e corrigirmos situações inesperadas como o maior erro do algoritmo para o sexo masculino.


8. REFERÊNCIAS


1. Gilli, G. et al. The Assessment of Skeletal Maturation. *Hormone Research*. 1996; 45(suppl 2): 49-52.
2. van Rijn, RR; Thodberg, HH. Bone age assessment: automated techniques coming of age?. *Acta Radiologica*. 2013; 54(9): 1024-9.
3. Chartrand, G. et al. Deep Learning: A Primer for Radiologists. *RadioGraphics*. 2017; 37: 2113–2131.
4. Larson, D.B. et al. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Radiology*. 2018; 287(1): 313-322.
5. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. *PLoS ONE* 14(7): e0220242. <https://doi.org/10.1371/journal.pone.0220242>
6. Halabi, SS. et al. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology*. 2019; 290:498–503

8. ANEXOS

Q Boneage_bot Cancel

GLOBAL SEARCH

 BoneAge_bot
@BoneAge_bot

< Chats BoneAge_bot bot 

What can this bot do?
Predict Bone Age based on xray image


Today

/start 15:41 ✓

I'm the Bone Age Deep Learning bot.
Please send me a hand-wrist x-ray!
15:41

Remember I'm learning to help you, but I don't make a real diagnosis, OK? Any questions, just send them here with your email for contact.
15:41


Mensagem encaminhada
De Felipe Kitamura

 16:03 ✓


Female age: 1 y 2 m +- 4 m 11 d (MAD)
16:03

Male age: 1 y 8 m +- 4 m 11 d (MAD)
16:03

I wanna show you how I see this X-ray.
Wait a moment, please...
16:03

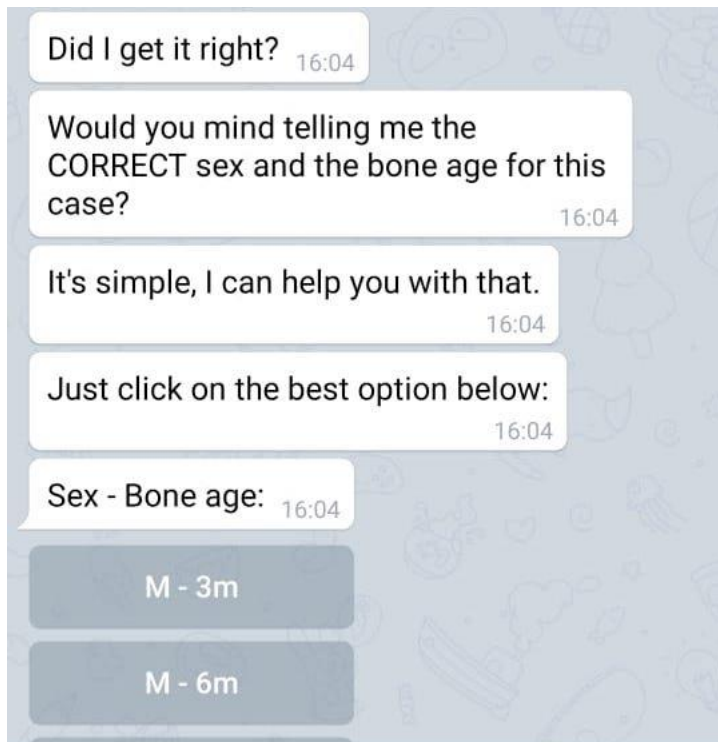
Artificial Intelligence Analysis
Activation map
 16:03
Warning: not for diagnostic purpose


Let me think... 16:03

 16:03

Felipe, it looks like... 16:03

The red spots were more important for my final decision.
16:04



😊 Mensagem  



😊 | Mensagem  