

Triagem automática de imagens de raio-x utilizando autoencoder: um método de validação de imagens médicas de mão e punho para a estimativa de idade óssea

Filipe Verrone de Lima

Programa de Pós Graduação

em Engenharia Elétrica e Computação
Universidade Presbiteriana Mackenzie

São Paulo, Brasil

verronefilipe@gmail.com

Rafael do Amaral Teodoro

Programa de Pós Graduação

em Engenharia Elétrica e Computação
Universidade Presbiteriana Mackenzie

São Paulo, Brasil

rafael.spteo@gmail.com

Gustavo Scalabrini Sampaio

Faculdade de Computação e Informática

Universidade Presbiteriana Mackenzie
São Paulo, Brasil

gustavo.sampaio@mackenzie.br

Cristiane Kochic

Faculdade de Ciências Médicas

Santa Casa de São Paulo

São Paulo, Brasil

ckochi@outlook.com

Carlos Alberto Longui

Faculdade de Ciências Médicas

Santa Casa de São Paulo

São Paulo, Brasil

carloslongui@msn.com

Leandro A. Silva

Programa de Pós Graduação

em Engenharia Elétrica e Computação

Universidade Presbiteriana Mackenzie

São Paulo, Brasil

leandroaugusto.silva@mackenzie.br

Resumo—Este trabalho apresenta um método de validação de imagens radiográficas de mão e punho com o fim de integrá-lo a uma aplicação clínica voltada à estimativa automatizada da idade óssea. A proposta consiste na definição de qualidade baseada em BAA e na construção de um modelo Convolutional Autoencoder (CAE) treinado exclusivamente com imagens de qualidade aceitável (classe positiva), a fim de permitir a classificação binária entre imagens válidas e inválidas segundo critérios objetivos de qualidade. Utilizou-se uma base original composta por 430 imagens de raio-X de mão e punho de pacientes saudáveis fornecida pela Santa Casa de Misericórdia de São Paulo. Essas imagens foram degradadas sinteticamente e controladamente em brilho e contraste, com variações graduais. O impacto dessas alterações foi mensurado pela variação na estimativa de idade óssea gerada por um modelo baseado em *Visual Geometry Group 16* (VGG16). Essa variação foi utilizada como base para definição dos limiares de qualidade aceitável. As imagens degradadas além desses limiares estatísticos foram chamadas de classe negativa. O modelo CAE foi treinado em uma estrutura de validação cruzada com cinco folds, avaliando diferentes arquiteturas. O modelo com melhor desempenho (CAE com cinco camadas) apresentou os menores erros de reconstrução na validação com a classe positiva entre os demais. O limiar de classificação foi definido utilizando a estatística *J* de Youden, priorizando a minimização da taxa de falso negativo, isto é, a não rejeição de imagens que de fato são válidas. O modelo obteve *recall* de 87,06% e *acurácia* de 63,88%.

Palavras-chave—qualidade de imagem médica, avaliação da idade óssea, radiografia de mão e punho, raio-X, Convolutional Autoencoder (CAE), *One-Class Classification* (OCC), *Machine Learning* em diagnóstico clínico

I. INTRODUÇÃO

A avaliação de qualidade de imagens (IQA, do inglês *Image Quality Assessment*) visa medir a qualidade de imagens

em relação à adequação para seu devido propósito [1], [2]. Essa avaliação pode ser feita por métodos subjetivos ou objetivos. Deste último método, são feitas três derivações, entre análise de referência total, parcial e ausente [1], [3]. A IQA é importante porque uma imagem inadequada pode levar à má interpretação, resultados e decisões incorretas ou sem fundamento, especialmente em contextos clínicos [4].

Em ferramentas médicas de exame automatizado, não é possível controlar os dados utilizados pelos usuários finais. Com isso, dados de baixa qualidade podem prejudicar o desempenho da ferramenta e elevar a incerteza dos resultados. Em aplicações de suporte à decisão clínica, isso representa um risco direto ao paciente, podendo resultar em diagnósticos incorretos e tratamentos desnecessários ou tardios [4], [5].

Entre os exames que utilizam imagens médicas e dependem de imagens com qualidade está a estimativa de idade óssea (BAA, do inglês *Bone Age Assessment*). O BAA é um exame realizado por pediatras, ortopedistas e endocrinologistas para auxiliar no monitoramento do desenvolvimento ósseo e no diagnóstico de distúrbios metabólicos e endócrinos [6]–[8]. O método de Greulich e Pyle (GP) é amplamente utilizado para BAA e consiste na comparação visual entre uma radiografia de mão e punho e um atlas de imagens de referência; a imagem do atlas que mais se assemelha à imagem do paciente é usada para atribuir a idade óssea do paciente [9]. Contudo, este método é subjetivo, depende da percepção do avaliador e pode apresentar uma variabilidade entre especialistas de até 1,5 anos na idade estimada [10], [11].

Outro método de BAA é o sistema de pontuação de Tanner-Whitehouse (TW). Esse método consiste na análise de partes

selecionadas da mão, como os ossos do carpo, ossos do metacarpo e falanges de dedos selecionados (a depender da variação de TW escolhida); cada parte é classificada em estágios de maturação entre A e I, à qual é convertida em uma pontuação. A estimativa final da idade óssea é obtida aplicando-se uma fórmula às pontuações resultantes das comparações. Ambas as metodologias costumam levar cerca de 30 minutos, de acordo com especialistas; são manuais e dependem da percepção visual individual de cada profissional [9], [12].

Visando possibilitar uma análise determinística e sem viés perceptual, foram propostos métodos alternativos utilizando *Machine Learning*. [11] propôs uma CNN (do inglês, *Convolutional Neural Network*) que utiliza cinco regiões anatômicas de interesse (ROI, do inglês *Regions Of Interest*), como carpos, grupos superior/inferior dos dedos, dedo mínimo e meia-mão, alcançando erros de 0,72 ano. Já [10] desenvolveu uma CNN hierárquica com 3 estágios com detecção de objetos usando 13 ROI's obtendo erro de aproximadamente 0,54 ano.

Portanto, há métodos computacionais com desempenho suficiente para desenvolvimento de aplicações de auxílio ao diagnóstico, porém há uma dependência da qualidade das imagens. Estudos demonstram que imagens com ruído (como o ruído Poisson) ou com brilho e contraste comprometidos afetam significativamente tanto a acurácia da avaliação humana quanto o desempenho de modelos computacionais de apoio à tomada de decisão [13], [14].

Este trabalho tem por objetivo propor um método de validação de imagens de raio-X de mão e punho para identificar imagens inadequadas para um modelo de estimativa de idade óssea. Para isso, utilizou-se um modelo Autoencoder Convolucional (CAE, do inglês *Convolutional Autoencoder*) treinado com *One-Class Classification* (OCC), a fim de classificar as imagens em duas categorias: com qualidade suficiente ou insuficiente. A proposta é incluir esse modelo como etapa de triagem em um aplicativo de estimativa de idade óssea.

O artigo foi estruturado da seguinte forma: a Seção II traz uma explicação da metodologia proposta e os referenciais teóricos de cada etapa do processo; a Seção III apresenta os resultados obtidos; e a Seção IV resume as conclusões e propõe trabalhos futuros.

II. METODOLOGIA

A. Base de Dados utilizada na Pesquisa

Criado pela Santa Casa de Misericórdia de São Paulo, o banco de dados utilizado neste artigo é original e contém dados de 430 pacientes saudáveis da população brasileira. Cada paciente possui uma identificação única, sexo (masculino ou feminino), uma imagem de raio-X contendo mão e punho, e a idade óssea estimada por médicos usando o método Greulich e Pyle (GP) [9]. As imagens de raio-X podem apresentar resoluções diferentes e uma parte delas está representada na Figura 1.

A população da base de dados é composta por 52% de pacientes do sexo feminino. A idade mínima da população é de 6 anos e a máxima de 18 anos.

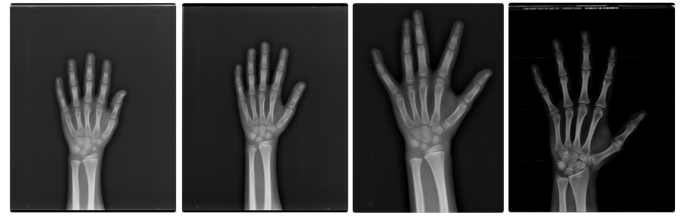


Figura 1. Imagens raio-X de mão e punho do banco de dados.

B. Arquitetura da Rede Neural Convolucional

Como forma de avaliar o impacto da qualidade das imagens em uma estimativa de idade óssea, foi utilizada uma rede neural artificial *Visual Geometry Group 16-layer* (VGG16) [15] modificada que processa o sexo do paciente e uma imagem raio-X contendo mão e punho – Figura 2. A imagem é decodificada nas camadas convolucionais, abstraindo informações importantes na estimativa da idade óssea, que posteriormente é processada junto ao sexo nas camadas densas. Por fim, o modelo gera uma saída contínua da idade relativa ao método GP. Esta rede foi escolhida e implementada por ser o estado da arte da literatura [10], [11]. O desempenho geral deste modelo não constitui o foco desta pesquisa, mas sim a estimativa gerada para uma determinada imagem, conforme será detalhado adiante. A premissa é que haverá variação na estimativa da idade óssea entre a imagem original da base de dados e sua versão deteriorada em termos de brilho ou contraste.

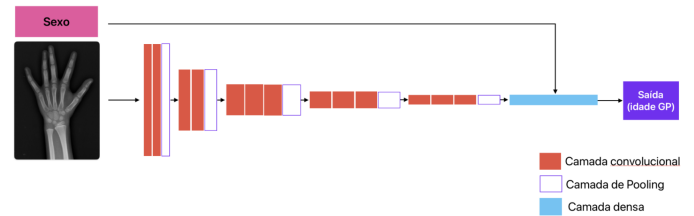


Figura 2. modelo baseado em VGG16 modificada.

C. Geração de imagens com variação de qualidade

De acordo com Chow et al. (2016) a avaliação de qualidade pode ser dividida em duas categorias, a saber: avaliação objetiva e subjetiva [3]. A avaliação objetiva divide-se em mais três categorias: referência total, reduzida e ausente, sendo que a referência se remete à uma imagem perfeita de referência para comparação com a imagem de teste. Neste trabalho, a qualidade é avaliada em análise objetiva de referência total, baseada na estimativa de idade óssea, utilizando como métrica a inferência feita pelo modelo da VGG16.

Primeiramente, é importante apresentar as definições sobre os dados utilizados neste trabalho. **Imagens originais:** São as imagens fornecidas pela Santa Casa, consideradas imagens com qualidade padrão desejável. **Imagens degradadas:** São as imagens originais degradadas em termos de características selecionadas. As modificações foram feitas em termos de brilho

e contraste, ambas características perceptuais importantes para clareza na análise [14]. Para cada característica selecionada, foi utilizada uma função de degradação que altera uma imagem em certa intensidade.

Para ambas as degradações, foi aplicada (1) [16]:

$$g(x) = \alpha f(x) + \beta \quad (1)$$

em que:

- $\alpha > 0$: é o fator conhecido como *ganho*, normalmente associado ao **contraste**;
- β : é o fator conhecido como *bias*, normalmente associado ao **brilho**.

As imagens originais foram degradadas em determinados fatores, definidos no intervalo $[-0,4; 0,45]$ com passo 0,01. Para o contraste, especificamente, o fator α foi definido no intervalo $[0,6; 1,45]$, pois o fator 1 não representa alteração. Cada imagem degradada passou pelo modelo de inferência e a idade óssea estimada foi salva.

Com esses dados, foi feita uma análise da variação de idade óssea (denominada *Bone Age Variation – BAV*) em anos, subtraindo a idade estimada de cada imagem degradada da imagem original correspondente, conforme (2):

$$B_v = B_d - B_o \quad (2)$$

em que:

- B_v : variação de idade óssea;
- B_d : estimativa da imagem degradada;
- B_o : estimativa da imagem original.

Os histogramas ilustrados na Figura 3 são as distribuições da BAV mencionada para cada função de degradação. A linha central vermelha indica a média e as linhas verde e azul estão posicionadas a dois desvios padrão para mais e para menos da média, respectivamente. Com base nos histogramas, foi definido o limiar de qualidade das radiografias, sendo de dois desvios padrão. A BAV aceita dentro desses limites não ultrapassa 0,36 ano (aproximadamente 4 meses e 10 dias). A definição de um limiar estatístico para BAV é essencial, pois a estimativa feita pelo modelo BAA incorpora não apenas os erros do próprio modelo, mas também os erros introduzidos por imagens de baixa qualidade. Assim, garante-se com esta proposta que apenas imagens dentro de um intervalo tolerável de BAV sejam aceitas, contribuindo para a redução do erro acumulado em um processo de estimativa de idade óssea em um aplicativo de auxílio ao diagnóstico.

Para cada característica degradada, foram agrupados os resultados por fator e verificados quais possuem média de BAV dentro dos limites estabelecidos e com desvio padrão igual ou menor que o desvio padrão global.

D. Classificação de qualidade da imagem

Esta é a etapa focal do trabalho. A proposta é desenvolver um modelo de redes neurais que receba uma radiografia de mão e punho e tenha como resultado um decisor, discriminando a imagem entre “válida” e “inválida”. Considera-se uma

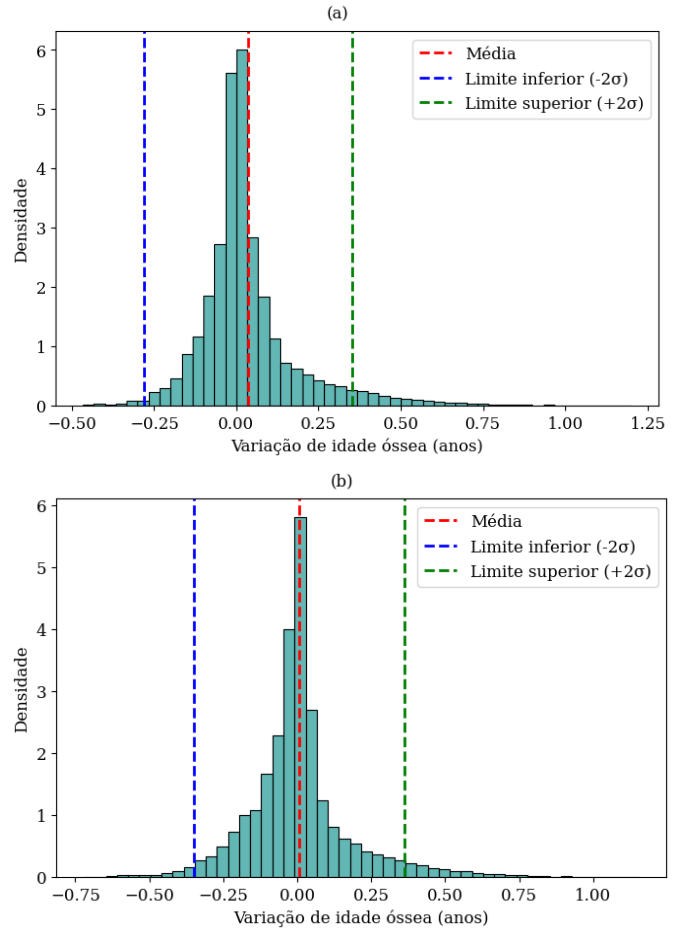


Figura 3. Distribuição de frequência das variações de idade óssea medidas entre imagens degradadas em brilho (a) e em contraste (b), comparadas às imagens originais (em anos).

imagem de entrada como “válida” se ela possui qualidade suficiente com base em similaridade entre a mesma e as imagens de treinamento do modelo, considerando as imagens fornecidas pela Santa Casa com qualidade suficiente para uso em diagnóstico.

Futuramente, se pensando em uma aplicação, a proposta é integrar esse modelo como uma etapa de triagem automática de um aplicativo, após a detecção de objetos e antes da inferência de idade óssea. A imagem cuja classe for identificada com grau de certeza suficiente na detecção de objeto passa pelo classificador de garantia de qualidade de raio-X. Se for considerada válida, segue o fluxo da aplicação até a inferência de idade óssea; caso contrário, há uma mensagem de retorno para o usuário.

Neste trabalho, foram consideradas imagens de classe negativa aquelas degradadas em brilho e contraste com fatores que ultrapassam os limites estatísticos definidos na seção II-C. Mas no uso clínico, a classe negativa não se limitaria a isso. Poderiam ocorrer imagens com qualidade suficiente nos termos das características avaliadas neste trabalho, porém com oclusão, ossos com fratura ou doença, como displasia. As

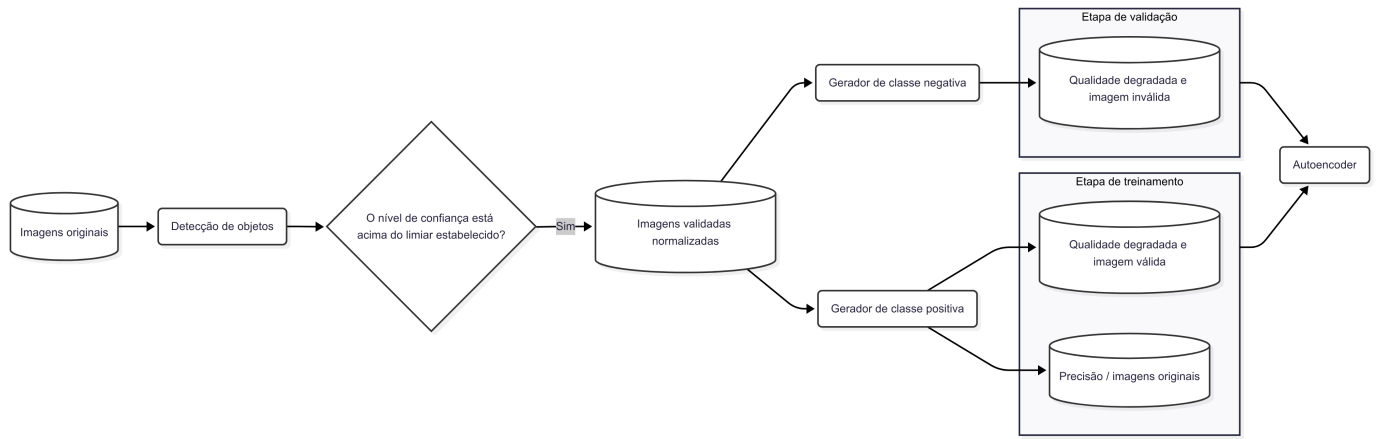


Figura 4. Experimento de preparação de imagens para treinamento e validação do CAE.

possibilidades para a criação de uma base de dados de classe negativa com as categorias listadas são numerosas e causariam evidente desbalanceamento dos dados na classificação. Por isso, foi considerado o uso de *One-Class Classification* (OCC), uma abordagem de treinamento não supervisionado que se mostrou eficiente em trabalhos similares anteriores ao utilizar no treinamento dados de uma classe somente, sem a necessidade de dados etiquetados [17], [18].

Entre as abordagens mais utilizadas dentro do OCC estão o deep SVDD (*deep Support Vector Data Description*, versão melhorada do SVDD adicionando um modelo de redes neurais), com o objetivo de modelar a distribuição dos dados normais e encontrar a menor região possível em volta dos pontos de dados, minimizada pela rede neural, resultando em um tipo de hiperplano para classificação [17]. Há também métodos baseados em reconstrução com modelos derivados do Autoencoder, que aplicam a detecção de anomalias de imagem de maneira não supervisionada. Espera-se que o erro de reconstrução de imagens anômalas seja maior que o erro de reconstrução de imagens normais. Entretanto, há trabalhos que demonstram que, se a base de dados de treinamento for diversa o suficiente, o modelo pode reconstruir bem, inclusive, imagens fora da classe esperada [17], [18].

O problema de classificação binária mencionado acima é abordado como um problema de detecção de anomalia. O modelo escolhido para as primeiras avaliações foi o Autoencoder Convolucional (CAE, do inglês *Convolutional Autoencoder*). O CAE é uma rede neural não supervisionada, projetada para aprender representações dos dados por meio de processos de codificação e decodificação utilizando camadas convolucionais. A hipótese é que o modelo treinado com a classe positiva seja capaz de reconstruir imagens desse conjunto com um erro menor que o erro de reconstrução de imagens da classe negativa. Dessa maneira, não é necessário usar extração de características para o treinamento, pois o CAE utiliza apenas a imagem de entrada para gerar uma representação latente que propõe manter somente características relevantes. Já no *decoder*, ele utiliza a representação latente para reconstruir a

imagem de entrada [19], [20].

E. Criação de uma base de dados para validação do One-class classification

O processo de criação da base de dados para treino e validação do modelo de OCC é apresentado na Figura 4. Foi definida como base de dados de treinamento inicial as 430 imagens fornecidas pela Santa Casa após passarem pela detecção de objetos, todas já padronizadas com 512x512 pixels, com a região da mão e punho recortada, 3 canais (3 cópias de um único canal em escala de cinza) e sempre na mesma direção vertical (porém, possivelmente em ambos os sentidos). Um exemplo de imagem padronizada pode ser visto na Figura 5. As imagens foram normalizadas com intensidade entre 0 e 1 e padronizadas com 256x256 pixels e 1 único canal em escala de cinza, ordenadas aleatoriamente e separadas em 5 folds para validação cruzada, cada fold com 86 imagens. Em cada fold, para cada imagem, foram geradas degradações utilizando os fatores considerados aceitáveis na análise estatística, sendo:

- 1) Brilho: intervalo de $-0,25$ a $0,15$, com passo de $0,01$, totalizando 40 variações de brilho; e
- 2) Contraste: intervalo de $-0,26$ a $0,45$, passo de $0,01$, totalizando 71 variações.

Também foram aplicadas inversões verticais e horizontais para cada imagem original de cada fold, sendo 86 inversões na vertical (N_v) e 86 inversões na horizontal (N_h). Todas as imagens de treinamento compõem a chamada classe positiva do treinamento do OCC e o número total utilizado nessa etapa foi, para cada fold, de $N_o + N_v + N_h + (40 + 71) * N_o = 9804$, em que N_o representa o número de imagens originais por fold e $40 + 71$ é a quantidade de degradações aplicadas a cada imagem original em brilho e contraste, respectivamente. O total foi de 49020 imagens.

Para gerar a chamada classe negativa foram consideradas para este trabalho apenas as imagens degradadas em qualidade nos termos de brilho e contraste, cujos fatores de degradação estão além dos limites estabelecidos, a saber:

- 1) Brilho: intervalo de $-0,4$ a $-0,26$ e de $0,16$ a $0,45$, com passo de $0,01$, totalizando 45 variações de brilho; e



Figura 5. Exemplo de raio-X após recorte da região de interesse.

- 2) Contraste: intervalo de $-0,4$ a $-0,27$, passo de $0,01$, totalizando 13 variações.

Essas variações foram aplicadas a cada imagem original e a suas versões com inversão vertical e horizontal, totalizando 74820 imagens de validação da classe negativa.

F. Treinamento do modelo

O modelo CAE foi treinado primeiramente em uma arquitetura com 5 camadas, conforme Figura 6. Na entrada, o modelo requer uma imagem normalizada com 256×256 pixels e um canal. Em cada camada do *encoder* há 3 componentes: um Conv2D com número de filtros variado no treinamento e função de ativação *relu*; um BatchNormalization para normalização; e um MaxPooling para redução de dimensionalidade pela metade. Após duas camadas, há o *bottleneck* da rede, onde está a versão codificada da imagem de entrada. No *decoder* há três camadas, as duas primeiras com os seguintes componentes: Conv2DTranspose com função de ativação *relu* como etapa oposta à convolução, com número de camadas espelhando as camadas do encoder; e um BatchNormalization para normalização novamente. A última camada do *decoder* é uma camada convolucional Conv2D com função de ativação *sigmoid* e um único filtro. Na saída, obtém-se a imagem reconstruída.

Para cada *fold*, as imagens da base de treinamento foram normalizadas e ajustadas às dimensões do modelo, adotando-se uma divisão de 90% para treino e 10% para validação. Foram utilizados dois *callbacks* durante o treinamento: *EarlyStopping*, com *restore_best_weights = True*, para interromper o treinamento caso a métrica de validação monitorada não apresentasse melhoria após um número definido de épocas, restaurando automaticamente os melhores pesos; e *ReduceLROnPlateau*, que reduz multiplicativamente a taxa de aprendizado quando a métrica monitorada estagna por certo número de épocas (*patience*).

Para avaliação do erro de reconstrução, foi utilizada a métrica erro absoluto médio (MAE, do inglês *Mean Absolute Error*). Dada uma imagem, a equação que define a métrica MAE é

$$\text{MAE} = \frac{1}{n} \sum_{k=0}^n |x_k - x'_k| \quad (3)$$

em que:

- n é o número total de pixels de uma imagem;
- x_k é o pixel da imagem de entrada;
- x'_k é o pixel da imagem reconstruída.

Também foram feitos testes utilizando um modelo com 7 camadas, adicionando uma camada convolucional no *encoder* e uma camada deconvolucional no *decoder*, também com número de filtros variável. A tabela I mostra os parâmetros utilizados no treinamento de cada variação do CAE proposto. Os demais parâmetros foram mantidos constantes, como é demonstrado a seguir:

- 1) Nas camadas convolucionais (convencional e transposta), foram utilizados filtros com *kernels* de convolução de tamanho 3×3 , com função de ativação *relu* (exceto a camada de saída, que utiliza função *sigmoid*) e zero *padding*, mantendo a mesma dimensão espacial da imagem de entrada;
- 2) MaxPooling com matriz 2×2 para redução da resolução espacial da imagem pela metade;
- 3) Nas camadas convolucionais transpostas foi realizado *upsampling* duplicando as dimensões espaciais da imagem; e
- 4) Na compilação do modelo foram utilizados o otimizador *adam* e função de perda MAE.

TABELA I
PARÂMETROS UTILIZADOS NO TREINAMENTO DO MODELO CAE.

Modelo	Nº de camadas e filtros
A	(128, 64)
B	(64, 32)
C	(32, 16)
D	(256, 128, 64)

III. RESULTADOS E DISCUSSÃO

Após o treinamento de cada variação do modelo, os modelos resultantes foram submetidos a etapas de validação, primeiramente com a classe positiva complementar do fold de treinamento, (por exemplo, se o modelo foi treinado no fold 1, a validação foi feita com os folds 2-5). A seguir, foi realizada a validação com a classe negativa. Para a análise dos resultados, os erros medidos conforme a subseção II-F foram salvos. Com base na métrica MAE de cada modelo na validação com classe positiva, conforme pode ser observado na tabela II, foi selecionado aquele com menor erro de reconstrução, a saber, o modelo A, com menor MAE. Nesta seção, são discutidas métricas e resultados deste modelo.

Dado que a proposta é integrar o modelo final em um aplicativo de processamento de imagem como uma etapa

Autoencoder de 5 camadas

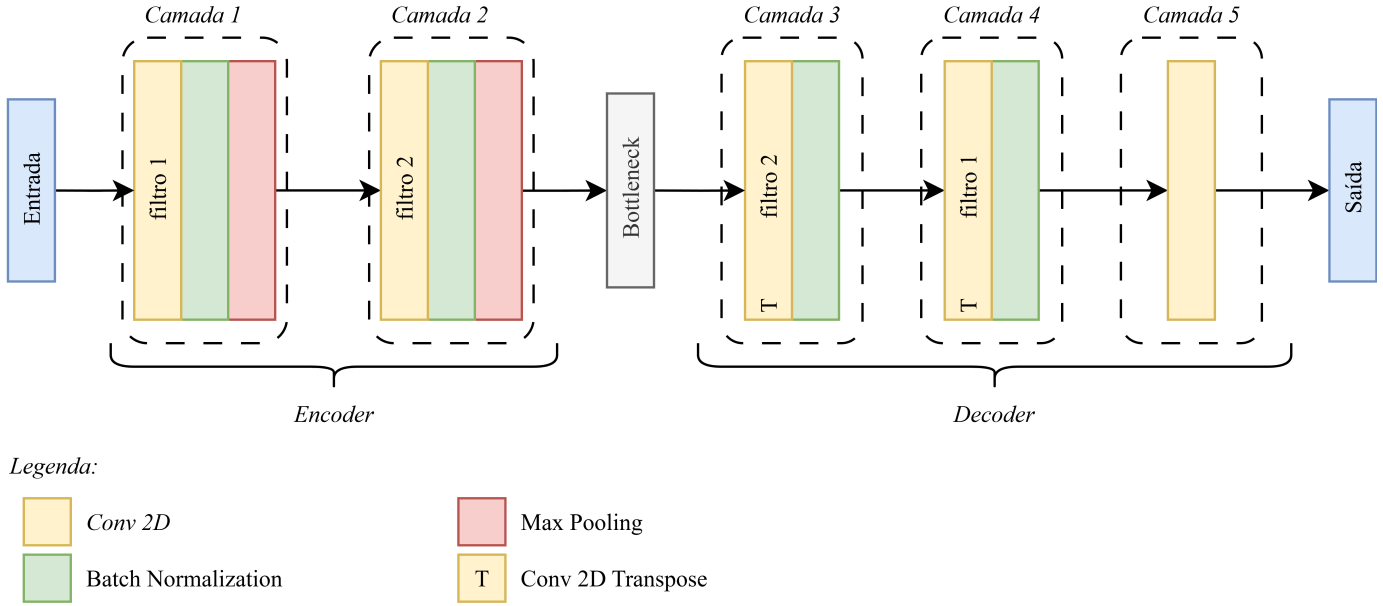


Figura 6. Arquitetura do Autoencoder Convolucional com 5 camadas.

TABELA II
DESEMPENHO DE CADA MODELO EM ETAPA DE VALIDAÇÃO COM A CLASSE POSITIVA

Modelo	MAE $\pm$ std (%)
A	0,5053 $\pm$ 0,1465
B	0,6773 $\pm$ 0,1825
C	0,8540 $\pm$ 0,2572
D	0,8288 $\pm$ 0,2203

de validação, o limiar de classificação foi otimizado para minimizar a taxa de falso negativo (FNR, do inglês *False Negative Rate*), equivalente ao complemento da métrica *recall* já que $FNR = 1 - Recall$. Dessa forma, foi priorizada a aceitação de entradas válidas, para não prejudicar a experiência do usuário ao negar imagens de raio-X potencialmente válidas. O limiar foi otimizado utilizando a estatística J de Youden [21] ($J = \text{sensitividade} + \text{especificidade} - 1$), procurando um ponto de equilíbrio entre as métricas de *especificidade* e de *sensitividade*, priorizando a redução de FNR. O limiar otimizado calculado foi de 0,00364. As métricas de classificação calculadas neste limiar, tais como *recall*, *especificidade* e *acurácia* são apresentadas na tabela III. A curva ROC (do inglês, *Receiving Operating Curve*) e AUC (do inglês, *Area Under Curve*) são apresentadas na Figura 7.

De acordo com as métricas apresentadas — matriz de confusão (Figura 8) e curva ROC e AUC (Figura 7) — observou-se que, em média, as imagens da classe negativa apresentaram erro de reconstrução inferior ao das imagens da classe positiva. Apesar de esse comportamento divergir do esperado em redes *autoencoder* (nas quais erros elevados cos-

TABELA III
DESEMPENHO MÉDIO E DESVIO PADRÃO DOS PRINCIPAIS INDICADORES DE AVALIAÇÃO DO MODELO.

Métrica	Valor Médio (%)	Desvio Padrão (%)
Acurácia	63,88	16,60
Precisão	51,44	9,83
Revocação (Recall)	87,06	9,75
Especificidade	51,73	29,59
F1-score	63,60	7,00
MAE (classe positiva)	0,5053	0,1465
MAE (classe negativa)	0,4343	0,1745

tumam indicar anomalias), ele pode refletir particularidades do conjunto de dados ou limitações no processo de modelagem, as quais devem ser investigadas.

Um dos fatores que possivelmente contribuíram para os resultados observados refere-se à diversidade da base de dados e à similaridade estrutural entre imagens de diferentes classes. A classe positiva, usada no treinamento, embora composta por imagens válidas, pode apresentar variações anatômicas significativas relacionadas à idade, sexo, posicionamento da mão e estágio de ossificação - fatores que podem afetar a capacidade de generalização do modelo. Considerando um cenário em que, a partir de uma mesma imagem original, são geradas duas versões degradadas — uma com variação de idade óssea dentro do limite aceitável e outra fora dele — é possível que ambas apresentem características estruturais muito semelhantes. Quando essas variações se situam próximas ao limiar estatístico de decisão (conforme definido na Seção II-C), a distinção entre classes tende a ser menos evidente.

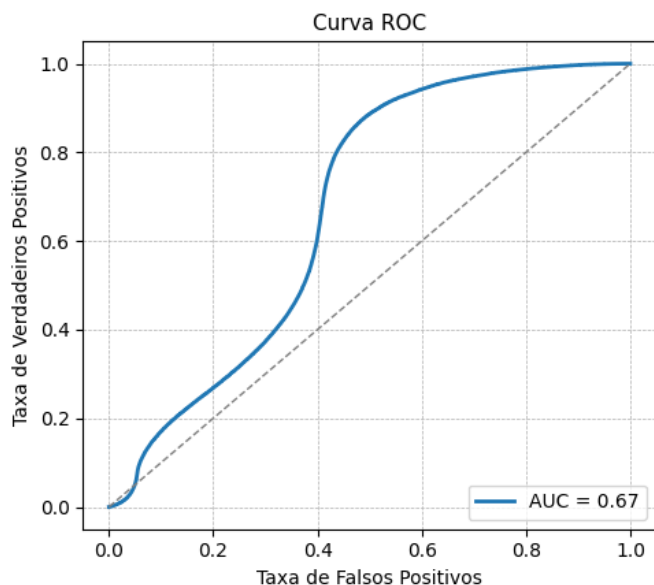


Figura 7. Curva ROC e melhor AUC encontrado

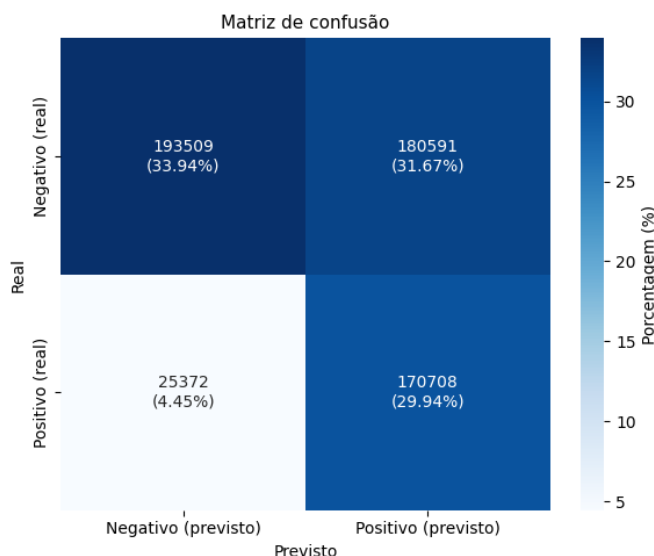


Figura 8. Matriz de confusão

IV. CONCLUSÃO E TRABALHOS FUTUROS

Este estudo propôs uma metodologia de validação de imagens radiográficas de mão e punho, com o objetivo de identificar automaticamente imagens inadequadas para aplicação em modelos de estimativa de idade óssea (BAA, do inglês *Bone Age Assessment*). O método foi concebido como uma etapa de triagem anterior à análise automatizada de BAA, a fim de assegurar a adequação das imagens utilizadas na inferência do modelo. A definição de imagem válida / adequada foi fundamentada em uma métrica objetiva de qualidade de imagem, estabelecida a partir de variações de estimativa de idade óssea (denominadas *Bone Age Variation* – BAV), a partir de imagens

originais. Essas imagens originais, cedidas pela Santa Casa, foram submetidas a alterações controladas em níveis de brilho e contraste, gerando versões degradadas. As imagens originais e suas versões degradadas foram processadas por um modelo automatizado de BAA baseado em VGG16, e as variações nas estimativas foram medidas. Com base na distribuição dessas variações (BAV), foram construídos histogramas a partir dos quais foram definidos limiares estatísticos. Esses limiares foram utilizados para agrupar os fatores de degradação de brilho e contraste de acordo com o BAV resultante. Imagens cuja degradação ultrapassava tais limiares foram classificadas como inadequadas, enquanto aquelas dentro dos limites estabelecidos foram consideradas válidas.

Para a etapa de avaliação automatizada da qualidade das imagens (IQA, do inglês *Image Quality Assessment*), foi empregado um modelo Autoencoder Convolutivo (CAE, do inglês *Convolutional Autoencoder*) com abordagem *One-Class Classification* (OCC), na qual o modelo é treinado exclusivamente com exemplares de uma classe. A classe positiva, usada no treinamento, foi composta por imagens válidas, distribuídas aleatoriamente em validação cruzada com a aplicação de aumento de dados. A classe negativa, utilizada apenas na validação do modelo após o treinamento, foi gerada a partir das imagens originais, submetendo-as a inversões horizontal e vertical, seguidas da aplicação de degradações com fatores superiores aos limiares estatísticos previamente definidos.

Os resultados indicam um desempenho razoável em termos de *recall*, com média de $87,06\% \pm 9,75\%$, alinhado ao objetivo da metodologia, que priorizou a redução da taxa de falsos negativos (FNR, do inglês *False Negative Rate*). A elevação da FNR comprometeria diretamente a credibilidade e a aplicabilidade clínica da solução, ao promover a exclusão de imagens potencialmente válidas. As demais métricas avaliadas apresentaram desempenhos inferiores ao *recall*, possivelmente devido a fatores como a diversidade anatômica da base de dados e a semelhança estrutural entre imagens degradadas com fatores limítrofes.

Em suma, os resultados obtidos demonstram a viabilidade da triagem automática de imagens radiográficas por meio de reconstrução via CAE, embora ainda existam limitações que demandam aprimoramentos futuros.

Como trabalhos futuros, propõe-se a análise do erro de estimativa da idade óssea entre as imagens classificadas como válidas após a otimização da métrica *recall*. Tal análise permite quantificar o impacto cumulativo dos erros desde a aquisição da imagem até a estimativa pelo modelo, passando por todo o processo do aplicativo. Pretende-se também avaliar a distribuição dos erros de reconstrução nas classes de validação positiva e negativa, com o objetivo de identificar e interpretar *outliers*. Caso se verifique que o baixo erro de reconstrução médio das imagens da classe negativa decorre da existência de imagens similares às da classe positiva, será considerado um ajuste nos limiares estatísticos adotados. Ademais, são propostas melhorias na arquitetura do CAE, por meio de experimentação com o aumento da profundidade da rede, modificação de hiperparâmetros e abordagens complementa-

res. Espera-se, por fim, elevar a precisão do classificador de qualidade de imagem e promover maior capacidade de generalização do modelo para aplicações clínicas na estimativa automatizada da idade óssea.

REFERÊNCIAS

- [1] J. Yang, M. Lyu, Z. Qi, and Y. Shi, "Deep learning based image quality assessment: A survey," *Procedia Computer Science*, vol. 221, pp. 1000–1005, 2023. Tenth International Conference on Information Technology and Quantitative Management (ITQM 2023).
- [2] H. Robert, *Image Quality*, pp. 925–928. Cham: Springer International Publishing, 2023.
- [3] L. S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomedical Signal Processing and Control*, vol. 27, pp. 145–154, 5 2016.
- [4] R. Syed, R. Eden, T. Makasi, I. Chukwudi, A. Mamudu, M. Kamalpour, D. Kapugama Geeganage, S. Sadeghianasl, S. J. J. Leemans, K. Goel, R. Andrews, M. T. Wynn, A. ter Hofstede, and T. Myers, "Digital health data quality issues: Systematic review," *J Med Internet Res*, vol. 25, p. e42615, Mar 2023.
- [5] C. Weng, W. Lin, S. Dong, Q. Liu, and H. Zhang, "First, do no harm: Addressing ai's challenges with out-of-distribution data in medicine," *Clinical and Translational Science*, vol. 18, no. 1, p. e70132, 2025. e70132 CTS-2024-0373C.
- [6] A. L. Creo and W. F. Schwenk, "Bone age: a handy tool for pediatric providers," *Pediatrics*, vol. 140, no. 6, 2017.
- [7] J. M. Zerlin and R. J. Hernandez, "Approach to skeletal maturation," *Hand clinics*, vol. 7, no. 1, pp. 53–62, 1991.
- [8] V. Gilsanz and O. Ratib, *Hand bone age: a digital atlas of skeletal maturity*. Springer Science & Business Media, 2005.
- [9] W. W. Greulich and S. I. Pyle, "Radiographic atlas of skeletal development of the hand and wrist," *The American Journal of the Medical Sciences*, vol. 238, no. 3, p. 393, 1959.
- [10] X. Xu, H. Xu, and Z. Li, "Automated bone age assessment: A new three-stage assessment method from coarse to fine," *Healthcare (Switzerland)*, vol. 10, 11 2022.
- [11] J. Zhou, Z. Li, W. Zhi, B. Liang, D. Moses, and L. Dawes, "Using convolutional neural networks and transfer learning for bone age classification," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6, Nov 2017.
- [12] J. Tanner, M. Healy, H. Goldstein, and N. Cameron, "Assessment of skeletal maturity and prediction of adult height: Tw3 method saunders," *London, UK*, pp. 1–110, 2001.
- [13] D. N. Thanh, V. B. Prasath, and L. M. Hieu, "A review on ct and x-ray images denoising methods," 2019.
- [14] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," *Handbook of image and video processing*, vol. 110, 2000.
- [15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [16] R. Szeliski, *Computer Vision*. Springer London, 2011.
- [17] M. Kim, K.-R. Moon, and B.-D. Lee, "Unsupervised anomaly detection for posteroanterior chest x-rays using multiresolution patch-based self-supervised learning," *Scientific Reports*, vol. 13, p. 3415, 2 2023.
- [18] P. Perera and V. M. Patel, "Learning deep features for one-class classification," *IEEE Transactions on Image Processing*, vol. 28, pp. 5450–5463, 11 2019.
- [19] S. M. Akila, E. Imanov, and K. Almezghwi, "Investigating beta-variational convolutional autoencoders for the unsupervised classification of chest pneumonia," *Diagnostics*, vol. 13, p. 2199, 7 2023.
- [20] N. Seliya, A. A. Zadeh, and T. M. Khoshgoftaar, "A literature review on one-class classification and its potential applications in big data," *Journal of Big Data*, vol. 8, 12 2021.
- [21] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.