

Inteligência artificial aplicada à radiografia de tórax: comparação de modelos de redes neurais convolucionais na detecção de opacidades.

Artificial intelligence applied to chest radiography: comparison of convolutional neural network models for opacity detection.

Maria Rita Bessa Muniz¹

Eugênia Maria das Chagas Silva²

Rendrick Alexandre Alemão Rogério³

Pedro Henrique Corrêa Teixeira⁴

Ana Beatriz Rodrigues Calegario⁵

Vinicius dos Santos Albuquerque⁶

Carolina Augusta Modena Heming⁷

Carolina Pontes Soares⁸

Ronaldo Francisco Ribeiro Pereira⁹

Rita de Cássia Ribeiro Pereira¹⁰

RESUMO

Introdução: A inteligência artificial, especialmente por meio de redes neurais convolucionais (CNNs), tem demonstrado crescente aplicabilidade na análise automatizada de exames de imagem, podendo auxiliar na identificação de alterações em radiografias de tórax. **Objetivo:** Desenvolver e comparar o desempenho de diferentes arquiteturas de redes neurais convolucionais na classificação de radiografias de tórax nas categorias Normal, Opacidade e Outras alterações, com ênfase na identificação de

¹ Universidade Federal do Acre, Brasil. E-mail: maria.rita@sou.ufac.br

² Universidade Federal do Acre, Brasil. E-mail: eugeniachag@gmail.com

³ Centro Universitário Uninorte, Brasil. E-mail: rendrickalemao@gmail.com

⁴ Universidade Federal do Acre, Brasil. E-mail: pedro.teixeira@sou.ufac.br

⁵ Universidade Federal do Acre, Brasil. E-mail: biacalegario21@gmail.com

⁶ Universidade Federal do Acre, Brasil. E-mail: Vnc.albq@gmail.com

⁷ Universidade de Iowa, Estados Unidos (EUA). E-mail: carolina-heming@uiowa.edu

⁸ Universidade Federal do Acre, Brasil. E-mail: carolina.soares@ufac.br

⁹ Universidade Federal do Acre, Brasil. E-mail: ronaldoffc_99@hotmail.com

¹⁰ Universidade Federal do Acre, Brasil. E-mail: ritapereira@ufac.br - Contato: ritapereira@ufac.br

opacidades. **Metodologia:** Trata-se de um estudo retrospectivo, analítico e observacional, no qual foram analisadas 3.349 radiografias de tórax de pacientes com idade igual ou superior a 18 anos. Foram avaliadas três arquiteturas: EfficientNetB0 e MobileNetV2, utilizando *transfer learning*, e uma CNN desenvolvida e treinada do zero. O desempenho dos modelos foi avaliado por meio de acurácia, precisão, sensibilidade, F1-score, área sob a curva ROC (AUC) e matrizes de confusão. **Resultados:** A EfficientNetB0 apresentou o melhor desempenho global, com acurácia de 85,9% e AUC macro próxima de 0,96. Especificamente para a identificação de opacidades, apresentou precisão de 84,6%, sensibilidade de 82,2% e F1-score de 83,4%. A CNN desenvolvida do zero apresentou acurácia global de 72,3%, com precisão de 83,5%, sensibilidade de 57,6% e F1-score de 68,2% para opacidades, além de indícios de *overfitting*. A MobileNetV2 apresentou desempenho global inferior, com acurácia próxima de 67%. **Conclusão:** A EfficientNetB0 associada ao *transfer learning* apresentou desempenho superior às demais arquiteturas avaliadas, especialmente na identificação de opacidades em radiografias de tórax. Os resultados demonstram o potencial das CNNs como ferramentas auxiliares à análise radiológica; entretanto, são necessários estudos com validação externa, maior diversidade de imagens e aprimoramento da interpretabilidade antes de sua aplicação clínica.

Palavras-chave: Inteligência artificial; Radiografia de tórax; Redes neurais convolucionais; *Deep learning*; Opacidade pulmonar.

ABSTRACT

Introduction: Artificial intelligence, particularly through convolutional neural networks (CNNs), has shown increasing applicability in the automated analysis of medical images and may assist in identifying abnormalities on chest radiographs. **Objective:** To develop and compare the performance of different convolutional neural network architectures for classifying chest radiographs into Normal, Opacity, and Other Abnormalities categories, with emphasis on opacity detection. **Methods:** This retrospective, analytical, observational study analyzed 3,349 chest radiographs from patients aged 18 years or older. Three architectures were evaluated: EfficientNetB0 and MobileNetV2 using transfer learning, and a CNN developed and trained from scratch. Model performance was assessed using accuracy, precision, sensitivity (recall), F1-score, area under the receiver operating characteristic curve (AUC), and confusion matrices. **Results:** EfficientNetB0 achieved the best overall performance, with an accuracy of 85.9% and a macro AUC of approximately 0.96. Specifically for opacity detection, it achieved 84.6%

precision, 82.2% sensitivity, and an F1-score of 83.4%. The CNN trained from scratch achieved an overall accuracy of 72.3%, with 83.5% precision, 57.6% sensitivity, and an F1-score of 68.2% for opacity detection, as well as evidence of overfitting. MobileNetV2 showed lower overall performance, with an accuracy of approximately 67%. Conclusion: EfficientNetB0 combined with transfer learning outperformed the other evaluated architectures, particularly in detecting opacities on chest radiographs. These findings demonstrate the potential of CNNs as auxiliary tools for radiographic analysis; however, external validation, greater image diversity, and improved model interpretability are required before clinical implementation.

Keywords: Artificial intelligence; Chest radiography; Convolutional neural networks; Deep learning; Pulmonary opacity.

INTRODUÇÃO

Vivenciamos uma nova revolução industrial, impulsionada pelo avanço de tecnologias como a Inteligência Artificial (IA), que permite às máquinas executarem não apenas tarefas manuais, mas também atividades racionais associadas à inteligência humana. Um exemplo marcante é o desenvolvimento de veículos autônomos, como os da Google e da Tesla, cujo uso plenamente independente ainda depende de legislação específica e de testes adicionais em cenários reais e de alto risco, exigindo sistemas robustos e seguros. Outra área de destaque é a dos diagnósticos automáticos, que já alcançam níveis de precisão comparáveis ou superiores aos de profissionais de saúde. A empresa iFlytek, por exemplo, desenvolveu um robô aprovado no exame nacional de licenciamento médico da China, capaz de registrar sintomas, analisar exames de imagem e realizar diagnósticos iniciais, com o objetivo de auxiliar médicos e aumentar sua eficiência, e não de substituí-los (1).

A radiologia destaca-se como uma das áreas médicas com maior aplicação da inteligência artificial. As radiografias de tórax figuram entre os exames de imagem mais realizados em âmbito mundial, totalizando aproximadamente 2 bilhões de procedimentos por ano. Nesse contexto, as redes neurais convolucionais profundas (Deep CNNs) têm demonstrado desempenho diagnóstico comparável ao de radiologistas humanos, conforme evidenciado em estudos que confrontam diretamente os resultados emitidos por sistemas de IA com aqueles produzidos por especialistas da área (5).

O uso da Inteligência Artificial baseia-se na aplicação de redes neurais artificiais (RNAs), formadas por neurônios interligados, em que individualmente são capazes de processar um dado e apresentar um resultado (3). Quando esses mecanismos são aplicados no ramo radiológico, é utilizada a Rede Neural Convolucional (CNN) que é um tipo específico de RNA, capaz de filtrar e distinguir a diferença existente entre duas imagens. Para o funcionamento dessa Rede Neural Convolucional, faz-se necessário a aplicação de um aprendizado por experiência, funcionando como uma fonte de "alimentação" para o sistema utilizado. Isso é feito por meio do Aprendizado de Máquina (AM), sendo capaz, portanto, de reconhecer padrões em exames de imagem (4).

A arquitetura do deep learning, conhecida também por aprendizado profundo, consiste em uma pilha de multicamadas de módulos simples, estando todos, ou a maioria, sujeitos ao aprendizado, capazes de computar mapeamentos de entrada-saída não lineares. Cada módulo na pilha transforma sua entrada para aumentar a seletividade e invariância da representação. Assim, um sistema com múltiplas camadas não lineares é capaz de ser simultaneamente sensível a detalhes mínimos, distinguindo, por exemplo, samoiedas de lobos brancos, e insensíveis a grandes variações irrelevantes, como o fundo, pose, iluminação e objetos ao redor (5).

Sendo assim, o funcionamento da IA consiste em analisar padrões de imagem, tendo como alicerce um banco de dados pré-existente proporcionando, então, avaliações quantitativas de aspectos da radiografia, contribuindo com as análises qualitativas feitas pelos profissionais, além de otimizar a qualidade dos diagnósticos e a velocidade do laudo do exame (2).

Este trabalho fundamenta-se na necessidade de compreender e avaliar a aplicação da inteligência artificial na análise automatizada de radiografias de tórax, especialmente na identificação de opacidades e outras alterações radiográficas. Nesse contexto, o objetivo deste estudo foi desenvolver e comparar o desempenho de diferentes arquiteturas de redes neurais convolucionais na classificação de radiografias de tórax nas categorias Normal, Opacidade e Outras alterações, com ênfase na identificação de opacidades. Foram avaliadas arquiteturas baseadas em transfer learning, EfficientNetB0 e MobileNetV2, além de uma rede neural convolucional desenvolvida e treinada do zero. A partir dos resultados obtidos, buscou-se analisar o desempenho das diferentes arquiteturas e discutir seu potencial como ferramentas

auxiliares à análise de radiografias de tórax.

MATERIAIS E MÉTODOS

Trata-se de um estudo retrospectivo, analítico e observacional, desenvolvido com o objetivo de avaliar e comparar o desempenho de modelos baseados em redes neurais convolucionais (CNNs) na classificação automatizada de radiografias de tórax em três categorias: Normal, Opacidade e Outras alterações.

As radiografias de tórax foram obtidas a partir de uma clínica de diagnóstico por imagem, por meio do sistema PACS (Picture Archiving and Communication System), e estavam associadas a laudos digitais elaborados por médicos especialistas em radiologia. O projeto foi submetido e aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal do Acre (CEP-UFAC), sob o parecer nº 6.186.816 e CAAE nº 55422622.5.0000.5010, com dispensa do Termo de Consentimento Livre e Esclarecido, conforme a legislação vigente.

Foram incluídos exames de pacientes com idade igual ou superior a 18 anos, realizados no período de julho a dezembro de 2022. Foram excluídos exames provenientes de indivíduos pertencentes a comunidades indígenas, pacientes legalmente incapazes, registros com dados incompletos e imagens com artefatos ou qualidade técnica inadequada, decorrentes de fatores como erros de posicionamento, movimentação ou inspiração insuficiente.

Após a aplicação dos critérios de inclusão e exclusão e a organização do banco de dados, foram incluídas 3349 radiografias, posteriormente distribuídas entre as classes Normal, Opacidade e Outras alterações.

Para o processamento das imagens, foram empregadas técnicas de pré-processamento destinadas à padronização dos dados de entrada. As radiografias foram redimensionadas de acordo com a resolução utilizada em cada experimento e tiveram os valores dos pixels normalizados para adequação ao treinamento das redes neurais.

A anonimização foi realizada diretamente no sistema PACS, permitindo a exportação das imagens no formato DICOM sem informações que possibilitaram a identificação dos pacientes.

GROUND TRUTH

O padrão de referência (ground truth) foi estabelecido com base nos laudos radiológicos associados às imagens, elaborados e assinados por médicos especialistas em radiologia vinculados à clínica de diagnóstico por imagem.

Inicialmente, os exames foram organizados de acordo com os achados descritos nos respectivos laudos radiológicos, incluindo categorias como Normal, Opacidade, COVID-19, pneumonia viral, alongamento aórtico, nódulo, aumento da área cardíaca, atelectasia, velamento dos seios costofrênicos, infiltrado intersticial e consolidação.

Para o desenvolvimento da tarefa de classificação, os exames foram posteriormente agrupados em três categorias finais: Normal, Opacidade e Outras alterações. A classe Normal compreendeu os exames sem alterações radiográficas; a classe Opacidade corresponde a áreas hipotransparência comparadas com o tecido pulmonar e adjacência ; e a classe Outras alterações reuniu alongamento aórtico, consolidação, infiltrado intersticial, velamento do seio costofrênico, atelectasia, área cardíaca aumentada, nódulo, Covid e pneumonia viral.

A etapa de coleta, organização e rotulagem dos dados foi conduzida a partir da análise dos laudos radiológicos e das respectivas imagens, sob supervisão de orientador e revisor com título de especialista em Radiologia e experiência profissional superior a dez anos. Posteriormente, os dados foram estruturados para processamento computacional utilizando a linguagem Python.

DATA PARTITIONS

A organização do conjunto de dados foi realizada considerando as limitações de hardware disponíveis e o custo computacional envolvido no treinamento dos modelos. As imagens radiográficas foram redimensionadas para três diferentes resoluções — 384×384, 256×256 e 128×128 pixels — às quais foram associados tamanhos de lote (batch size) de 8, 16 e 32 imagens, respectivamente. A escolha dessas resoluções demonstrou desempenho satisfatório nos experimentos realizados e esteve diretamente relacionada às restrições impostas pela infraestrutura computacional disponível, composta por um processador Intel Core i7-9700K, 64 GB de memória RAM e uma unidade de processamento gráfico NVIDIA RTX 3060 com 12 GB de memória dedicada.

A divisão do conjunto de dados foi estabelecida na proporção de 80% para treinamento e 20% para teste. Em função do número limitado de amostras, o subconjunto destinado ao teste foi igualmente empregado para a etapa de validação dos modelos.

Os metadados associados às imagens incluíram um identificador numérico do paciente, que não possibilita sua identificação, além de informações referentes ao sexo, idade e às incidências radiográficas disponíveis, compreendendo imagens de tórax em projeções pósterio-anterior e de perfil

MODEL

Foram avaliadas três arquiteturas baseadas em redes neurais convolucionais: EfficientNetB0, MobileNetV2 e uma CNN desenvolvida e treinada do zero.

As arquiteturas EfficientNetB0 e MobileNetV2 foram implementadas utilizando a estratégia de transfer learning, na qual modelos previamente treinados são adaptados à tarefa específica de classificação proposta neste estudo. Essa estratégia permite aproveitar representações visuais previamente aprendidas, reduzindo a necessidade de treinamento integral da rede a partir de pesos aleatórios.

Paralelamente, foi desenvolvida uma CNN do zero, permitindo comparar o desempenho de uma arquitetura treinada especificamente com o conjunto de dados do estudo com aquele obtido pelas arquiteturas pré-treinadas.

A camada de saída dos modelos foi configurada para realizar a classificação das imagens em três categorias: Normal, Opacidade e Outras alterações.

O ambiente de desenvolvimento utilizado foi o Anaconda Python, por meio do Jupyter Lab/Jupyter Notebook. Para implementação, processamento dos dados, treinamento e avaliação dos modelos, foram utilizadas bibliotecas como NumPy, Pandas, Keras e Scikit-learn.

TRAINING

Os modelos foram treinados utilizando os conjuntos de treinamento previamente definidos, com acompanhamento do desempenho no conjunto de validação. Durante o processo, foram empregadas técnicas de aumento de dados (data augmentation) com o objetivo de ampliar a variabilidade das imagens apresentadas às redes, incluindo

pequenas variações de contraste, rotações e inversões horizontais

Também foram empregadas estratégias de regularização e controle do treinamento, de acordo com as configurações estabelecidas para cada arquitetura, visando reduzir o risco de sobreajuste e favorecer a capacidade de generalização dos modelos.

Para cada época foram registradas a acurácia e a função de perda nos conjuntos de treinamento e validação. Na CNN desenvolvida do zero, cujas curvas de aprendizado são apresentadas nos resultados, foram analisadas cinco épocas de treinamento.

Os modelos finais foram avaliados no conjunto de testes por meio de métricas de desempenho e matrizes de confusão.

EVALUATION

O desempenho das arquiteturas foi avaliado no conjunto de teste por meio das métricas de acurácia, precisão, sensibilidade (recall), F1-score e área sob a curva ROC (AUC).

A acurácia foi utilizada para representar a proporção global de classificações corretas. A precisão expressou a proporção de predições de determinada classe que foram corretamente classificadas, enquanto a sensibilidade representou a capacidade do modelo de identificar corretamente as imagens pertencentes a cada classe. O F1-score correspondeu à média harmônica entre precisão e sensibilidade.

Considerando a natureza multiclasse da tarefa, foram calculadas métricas específicas para as classes Normal, Opacidade e Outras alterações, bem como médias macro e ponderadas. A média macro atribuiu o mesmo peso a cada classe, enquanto a média ponderada considerou a quantidade de amostras pertencentes a cada categoria.

Adicionalmente, foram construídas matrizes de confusão em valores absolutos e normalizados, permitindo identificar a distribuição dos acertos e erros entre as três classes. A área sob a curva ROC foi utilizada como medida complementar da capacidade discriminativa dos modelos.

A análise comparativa considerou o desempenho global das arquiteturas e, particularmente, sua capacidade de identificação da classe Opacidade.

$$Acurácia = \frac{VP+VN}{VP+VN+FP+FN} \quad (1)$$

$$Sensibilidade (recall) = \frac{VP}{VP+FN} \quad (2)$$

$$Precisão = \frac{VP}{VP+FP} \quad (3)$$

$$F1 = 2 \cdot \frac{Precisão \cdot recall}{Precisão + recall} \quad (4)$$

onde:

VP= verdadeiros positivos

VN= verdadeiros negativos

FP = falsos positivos

FN= falsos negativos

CONSIDERAÇÕES ÉTICAS E LIMITAÇÕES

Todas as imagens analisadas foram previamente anonimizadas, não contendo informações que permitissem a identificação dos indivíduos. O estudo foi desenvolvido em consonância com os preceitos éticos previstos na Resolução nº 466/2012 do Conselho Nacional de Saúde e aprovado pelo Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal do Acre (CEP-UFAC), sob parecer nº 6.186.816 e CAAE nº 55422622.5.0000.5010.

Entre as limitações metodológicas, destacam-se a utilização de dados provenientes de uma única fonte institucional, o desbalanceamento entre as classes e a ausência de validação externa em uma base independente. Além disso, o agrupamento de diferentes achados radiográficos na categoria Outras alterações pode aumentar a heterogeneidade dessa classe e dificultar sua discriminação em relação às demais categorias.

RESULTADOS

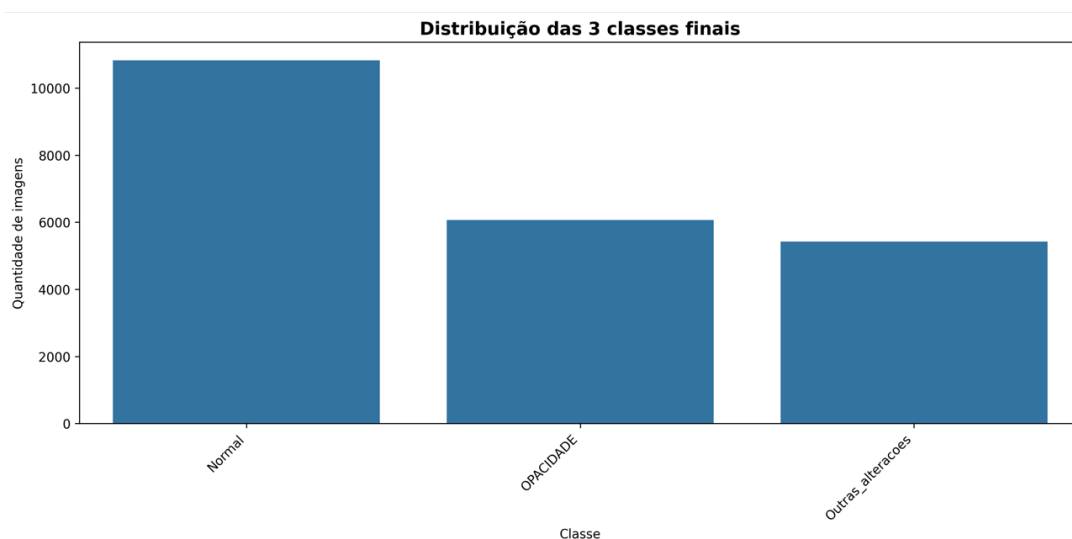
DISTRIBUIÇÃO DAS CLASSES

O conjunto de dados utilizado neste estudo apresentou inicialmente diferentes categorias de achados radiográficos, incluindo exames classificados como normais, opacidade, COVID-19, pneumonia viral, alongamento aórtico, nódulo, área cardíaca aumentada, atelectasia, velamento dos seios costofrênicos, infiltrado intersticial e consolidação.

Para o desenvolvimento dos modelos de classificação, os exames foram posteriormente agrupados em três classes finais: Normal, Opacidade e Outras alterações. Observou-se predominância da classe Normal, seguida pelas classes Opacidade e Outras alterações, caracterizando um desbalanceamento moderado entre as categorias. Apesar dessa diferença, as três classes permaneceram representadas nos conjuntos destinados ao treinamento, validação e teste.

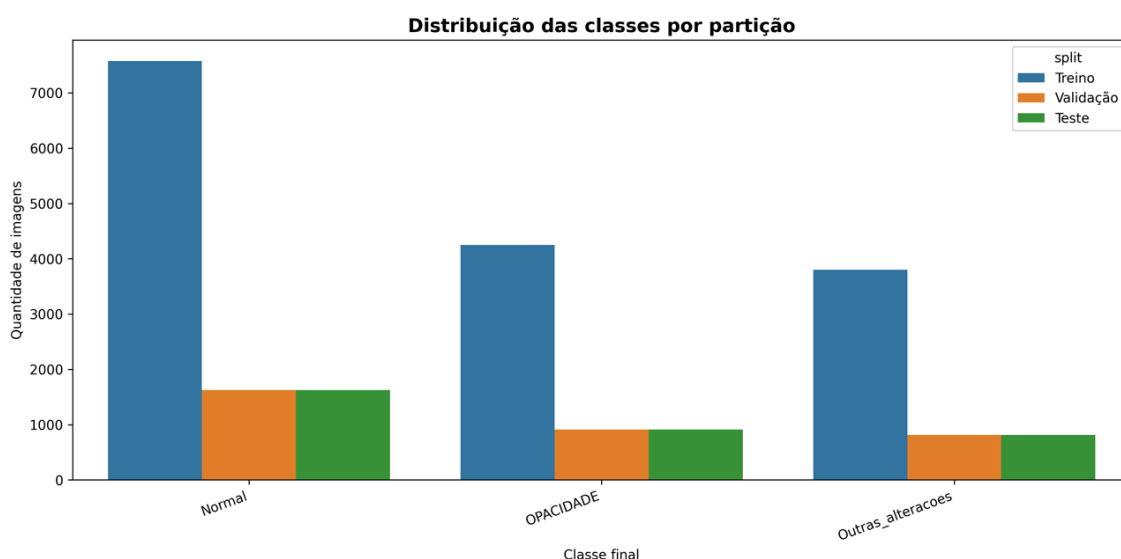
A distribuição entre as partições manteve proporções semelhantes para cada classe, com maior número de imagens destinado ao treinamento e parcelas menores destinadas à validação e ao teste. Essa estratégia permitiu preservar a representação das três categorias ao longo das diferentes etapas de desenvolvimento e avaliação dos modelos.

Figura 1. Distribuição das classes finais do conjunto de dados.



Fonte: Elaborado pelos autores.

Figura 2. Distribuição das classes entre os conjuntos de treinamento, validação e teste.



Fonte: Elaborado pelos autores.

DESEMPENHO DURANTE O TREINAMENTO

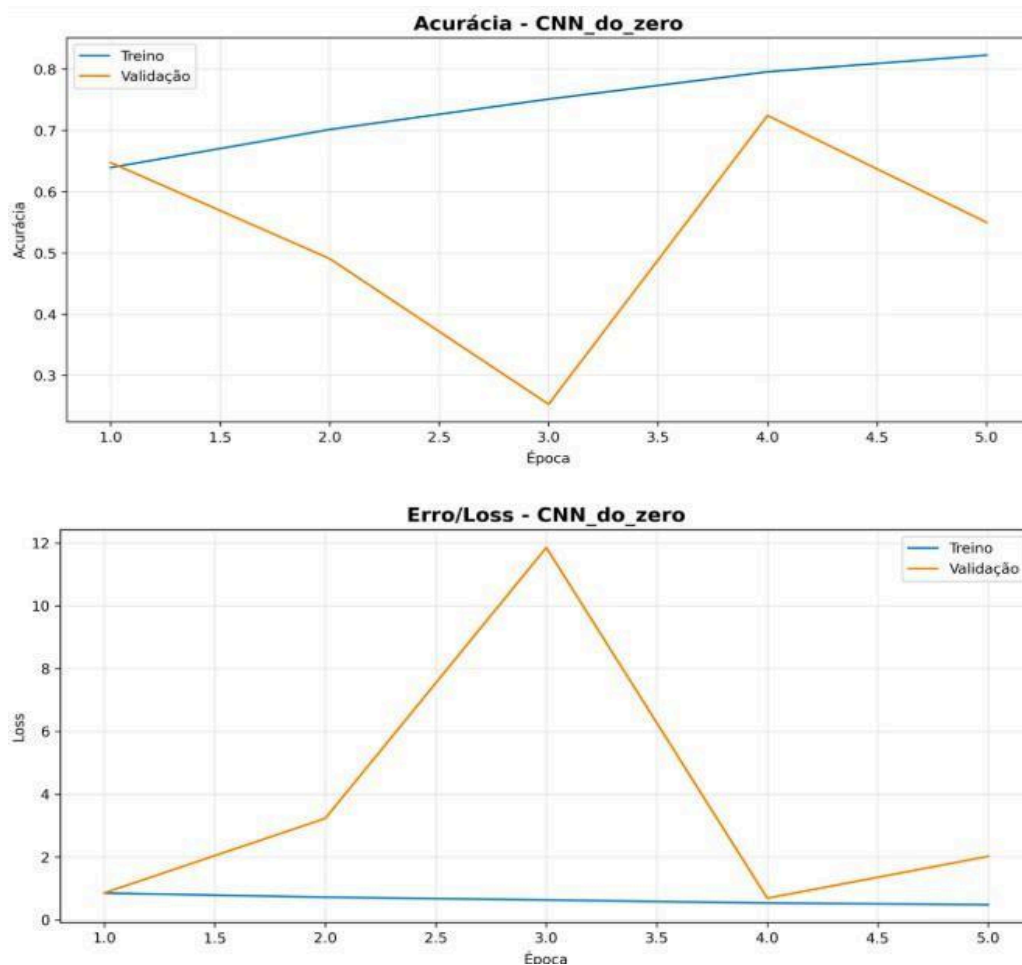
A evolução da acurácia e da função de perda (loss) da rede neural convolucional desenvolvida do zero (CNN_do_zero) durante cinco épocas é apresentada na Figura 3. Observou-se aumento progressivo da acurácia no conjunto de treinamento, partindo de aproximadamente 64% na primeira época e atingindo cerca de 82% na quinta época. Paralelamente, houve redução contínua da perda de treinamento, indicando progressivo ajuste do modelo aos dados utilizados nessa etapa.

Entretanto, comportamento semelhante não foi observado no conjunto de validação. A acurácia de validação apresentou oscilações expressivas entre as épocas, variando de aproximadamente 25% a 72%, com maior valor observado na quarta época e redução para aproximadamente 55% na quinta época. De maneira correspondente, a perda de validação apresentou elevada variabilidade, alcançando seu maior valor na terceira época e seu menor valor na quarta época.

A divergência entre a melhora contínua observada no treinamento e a instabilidade das métricas de validação indica menor capacidade de generalização da CNN desenvolvida do zero, compatível com ocorrência de sobreajuste (overfitting) durante o processo de treinamento.

Figura 3. Evolução da acurácia e da perda nos conjuntos de treinamento e validação da

CNN desenvolvida do zero.



Fonte: Elaborado pelos autores.

DESEMPENHO DOS MODELOS

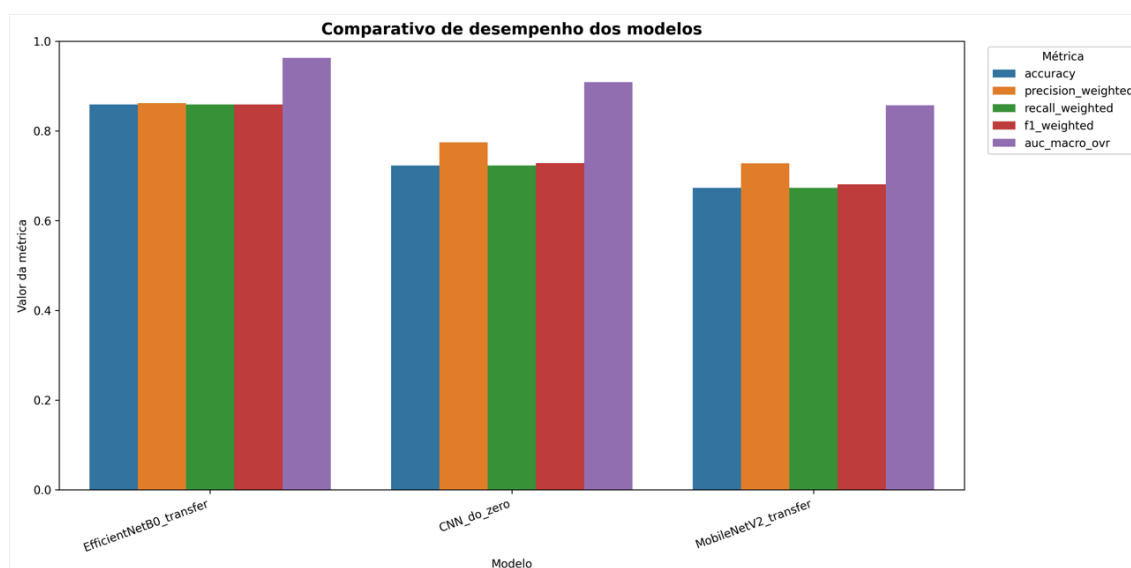
Foram avaliadas três estratégias de classificação: EfficientNetB0 com aprendizado por transferência (transfer learning), CNN desenvolvida do zero e MobileNetV2 com aprendizado por transferência. As arquiteturas apresentaram diferenças nas métricas de desempenho, conforme demonstrado na Figura 4.

Entre os modelos avaliados, a EfficientNetB0 apresentou acurácia global de aproximadamente 86%, enquanto a CNN desenvolvida do zero apresentou acurácia de aproximadamente 72%. A MobileNetV2 apresentou acurácia próxima de 67%. Considerando as demais métricas ponderadas apresentadas no comparativo, a EfficientNetB0 alcançou aproximadamente 86% de precisão, 86% de sensibilidade e 86% de F1-score, além de área sob a curva ROC (AUC) macro próxima de 0,96.

A CNN desenvolvida do zero apresentou aproximadamente 77% de precisão

ponderada, 72% de sensibilidade e 73% de F1-score, com AUC macro próxima de 0,91. Para a MobileNetV2, foram observados valores próximos de 73% para precisão, 67% para sensibilidade, 68% para F1-score e 0,86 para AUC macro.

Figura 4. Comparação das métricas de desempenho entre os modelos EfficientNetB0, CNN desenvolvida do zero e MobileNetV2.



Fonte: Elaborado pelos autores.

DESEMPENHO DA EFFICIENTNETB0

A matriz de confusão da EfficientNetB0 no conjunto de teste incluiu 3.349 imagens, sendo 1.624 pertencentes à classe Normal, 911 à classe Opacidade e 814 à classe Outras alterações.

Entre as 1.624 imagens normais, 1.382 (85%) foram corretamente classificadas, enquanto 113 (7%) foram classificadas como Opacidade e 129 (8%) como Outras alterações. Na classe Opacidade, 749 das 911 imagens (82%) foram corretamente identificadas, enquanto 107 (12%) foram classificadas como Normal e 55 (6%) como Outras alterações. Entre as 814 imagens pertencentes à classe Outras alterações, 746 (92%) foram corretamente classificadas, enquanto 45 (6%) foram classificadas como Normal e 23 (3%) como Opacidade.

A EfficientNetB0 apresentou, portanto, acurácia global de 85,9% (2.877/3.349). Especificamente para a classe Opacidade, foram obtidas precisão de 84,6%,

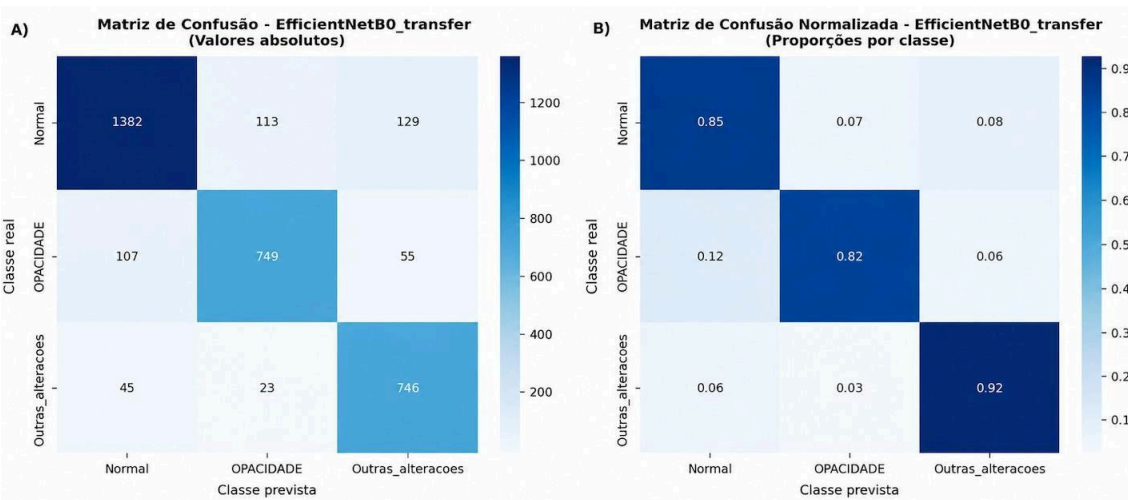
sensibilidade de 82,2% e F1-score de 83,4%.

Tabela 1. Métricas de desempenho da EfficientNetB0

Classe	Precisão	Sensibilidade	F1-score
Normal	0,90	0,85	0,88
Opacidade	0,85	0,82	0,83
Outras alterações	0,80	0,92	0,86
Média macro	0,85	0,86	0,85
Média ponderada	0,86	0,86	0,86

Fonte: Elaborado pelos autores.

Figura 5. Matriz de confusão da EfficientNetB0 no conjunto de teste.



Fonte: Elaborado pelos autores.

DESEMPENHO DA CNN DESENVOLVIDA DO ZERO

Na CNN desenvolvida do zero, das 1.624 imagens pertencentes à classe Normal, 1.165 (72%) foram corretamente classificadas. Entre as classificações incorretas, 80 (5%) foram atribuídas à classe Opacidade e 379 (23%) à classe Outras alterações.

Na classe Opacidade, 525 das 911 imagens (58%) foram corretamente identificadas, enquanto 130 (14%) foram classificadas como Normal e 256 (28%) como Outras alterações. Para a classe Outras alterações, 732 das 814 imagens (90%) foram

corretamente classificadas, enquanto 58 (7%) foram consideradas normais e 24 (3%) foram classificadas como Opacidade.

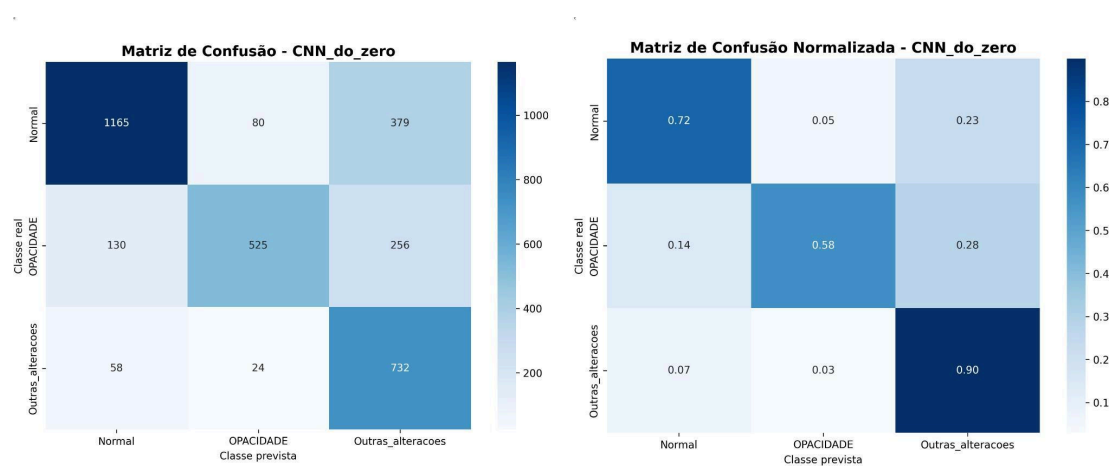
A acurácia global da CNN desenvolvida do zero foi de 72,3% (2.422/3.349). Para a classe Opacidade, o modelo apresentou precisão de 83,5%, sensibilidade de 57,6% e F1-score de 68,2%.

Tabela 2. Métricas de desempenho da CNN desenvolvida do zero

Classe	Precisão	Sensibilidade	F1-score
Normal	0,86	0,72	0,78
Opacidade	0,83	0,58	0,68
Outras alterações	0,54	0,90	0,67
Média macro	0,74	0,73	0,71
Média ponderada	0,77	0,72	0,73

Fonte: Elaborado pelos autores.

Figura 6. Matriz de confusão da CNN desenvolvida do zero no conjunto de testes.



Fonte: Elaborado pelos autores.

1.1 COMPARAÇÃO ENTRE AS ARQUITETURAS

A comparação entre as arquiteturas demonstrou diferenças principalmente na capacidade de identificação da classe Opacidade. A EfficientNetB0 classificou corretamente 82% das imagens com opacidade, enquanto a CNN desenvolvida do zero identificou corretamente 58% dos exames dessa classe.

Na CNN desenvolvida do zero, o principal erro envolvendo a classe Opacidade ocorreu na classificação como Outras alterações, correspondendo a 28% dos exames com opacidade (256/911). Na EfficientNetB0, essa proporção foi reduzida para 6% (55/911). A classificação de exames com opacidade como normais também foi menor na EfficientNetB0, correspondendo a 12% (107/911), em comparação com 14% (130/911) na CNN desenvolvida do zero.

A diferença também foi observada na acurácia global, que passou de 72,3% na CNN desenvolvida do zero para 85,9% na EfficientNetB0, correspondendo a um aumento absoluto de aproximadamente 13,6 pontos percentuais. A análise conjunta das métricas demonstra diferenças de desempenho entre as arquiteturas avaliadas, especialmente quanto à sensibilidade para identificação da classe Opacidade.

DISCUSSÃO

A radiografia de tórax é um dos exames de imagem mais utilizados na prática clínica, e o grande volume de imagens produzidas favorece a aplicação de técnicas de inteligência artificial para auxiliar sua interpretação. Nesse contexto, as redes neurais convolucionais (Convolutional Neural Networks – CNNs) apresentam capacidade de extrair automaticamente características das imagens e utilizá-las em tarefas de classificação, sendo amplamente estudadas na identificação de alterações em radiografias torácicas (6,7).

No presente estudo, foram avaliadas diferentes arquiteturas de redes neurais convolucionais para classificação de radiografias de tórax em três categorias: Normal, Opacidade e Outras alterações. Foram comparadas uma CNN desenvolvida e treinada do zero e arquiteturas baseadas em transfer learning, incluindo EfficientNetB0 e MobileNetV2. Entre os modelos avaliados, a EfficientNetB0 apresentou o melhor desempenho global, com acurácia de aproximadamente 86%, enquanto a CNN

desenvolvida do zero apresentou 72% e a MobileNetV2, aproximadamente 67%. A EfficientNetB0 também apresentou maiores valores globais de precisão, sensibilidade, F1-score e AUC.

O desempenho da EfficientNetB0 pode estar relacionado às vantagens do transfer learning. Essa estratégia utiliza representações previamente aprendidas em grandes conjuntos de imagens e posteriormente as adapta para uma nova tarefa, sendo especialmente útil quando a disponibilidade de dados médicos adequadamente rotulados é limitada (8,9). Estudos anteriores também demonstram a aplicabilidade de arquiteturas baseadas em deep learning, incluindo EfficientNet, para identificação automatizada de alterações em radiografias de tórax (10).

Na análise específica da classe Opacidade, a EfficientNetB0 apresentou sensibilidade de 82,2%, precisão de 84,6% e F1-score de 83,4%, enquanto a CNN desenvolvida do zero apresentou sensibilidade de 57,6%, precisão de 83,5% e F1-score de 68,2%. Embora os modelos tenham apresentado valores semelhantes de precisão para essa classe, a diferença na sensibilidade demonstra que a CNN desenvolvida do zero deixou de reconhecer uma proporção maior das imagens com opacidade.

Esse comportamento também foi evidenciado pelas matrizes de confusão. Das 911 imagens pertencentes à classe Opacidade, a EfficientNetB0 classificou corretamente 749, enquanto a CNN desenvolvida do zero identificou corretamente 525. Na EfficientNetB0, apenas 6% das imagens com opacidade foram classificadas como Outras alterações, enquanto na CNN desenvolvida do zero esse erro ocorreu em aproximadamente 28% dos casos. Esses resultados demonstram maior capacidade da EfficientNetB0 em diferenciar opacidades das demais alterações presentes nas radiografias.

Outro aspecto relevante foi o comportamento das curvas de treinamento da CNN desenvolvida do zero. A acurácia de treinamento aumentou progressivamente ao longo das cinco épocas, acompanhada pela redução da função de perda. Entretanto, as métricas de validação apresentaram oscilações consideráveis. Essa divergência entre melhora no conjunto de treinamento e instabilidade no conjunto de validação constitui possível indício de sobreajuste (overfitting), fenômeno no qual o modelo aprende

características específicas dos dados de treinamento, mas apresenta menor capacidade de generalização para dados não utilizados diretamente no ajuste dos parâmetros (7,11). Esse comportamento pode contribuir para explicar o desempenho inferior da CNN desenvolvida do zero em comparação à EfficientNetB0.

Por outro lado, os resultados não sugerem underfitting, uma vez que a CNN demonstrou capacidade de aprendizagem no conjunto de treinamento, evidenciada pelo aumento da acurácia e redução progressiva da perda. Assim, a principal limitação observada parece estar relacionada à capacidade de generalização do modelo, e não à incapacidade de aprender os padrões presentes nos dados de treinamento (11).

A distribuição das classes também deve ser considerada na interpretação dos resultados, uma vez que houve predominância de exames classificados como normais em relação às categorias Opacidade e Outras alterações. Em conjuntos de dados desbalanceados, a acurácia isoladamente pode não representar adequadamente o desempenho do modelo. Dessa forma, a utilização conjunta de precisão, sensibilidade, F1-score, AUC e matriz de confusão permite uma avaliação mais abrangente da capacidade de classificação, principalmente quando o desempenho varia entre as diferentes classes (12).

Apesar dos resultados obtidos, o desempenho apresentado no conjunto de teste não é suficiente para determinar a aplicabilidade clínica dos modelos. Algoritmos de inteligência artificial podem apresentar redução de desempenho quando avaliados em imagens provenientes de diferentes instituições, equipamentos, protocolos e populações. A validação externa é, portanto, fundamental para verificar a capacidade de generalização de um algoritmo antes de sua utilização em cenários clínicos distintos daqueles empregados durante seu desenvolvimento (13).

Entre as limitações deste estudo destaca-se, assim, a ausência de validação externa independente. Além disso, o agrupamento dos achados radiográficos em três categorias amplas pode reduzir a granularidade diagnóstica do modelo. A classe Outras alterações reúne diferentes achados radiográficos, enquanto a categoria Opacidade pode estar relacionada a diferentes condições, favorecendo a sobreposição de características entre as classes. Outra limitação inerente aos modelos de deep learning é sua reduzida

interpretabilidade, frequentemente denominada problema da “caixa-preta”. Métodos de inteligência artificial explicável, como mapas de ativação, podem contribuir para demonstrar quais regiões das imagens influenciam as decisões do algoritmo e aumentar a compreensão do processo de classificação (7,14).

Dessa forma, os resultados demonstram o potencial das redes neurais convolucionais para classificação automatizada de radiografias de tórax. Entre as arquiteturas avaliadas, a EfficientNetB0 com transfer learning apresentou melhor desempenho global, com acurácia de 85,9% e AUC macro próxima de 0,96, além de maior sensibilidade para identificação de opacidades quando comparada à CNN desenvolvida do zero. Entretanto, estudos adicionais com validação externa, maior diversidade de imagens e métodos de interpretabilidade são necessários para avaliar a capacidade de generalização e a possível aplicação desses modelos como ferramentas auxiliares na prática clínica.

CONCLUSÃO

Por meio do presente estudo, foi possível demonstrar o potencial da inteligência artificial, por meio de redes neurais convolucionais, para a classificação automatizada de radiografias de tórax nas categorias Normal, Opacidade e Outras alterações. Entre as arquiteturas avaliadas, a EfficientNetB0 com transfer learning apresentou o melhor desempenho, alcançando acurácia global de 85,9% e AUC macro próxima de 0,96, além de precisão de 84,6%, sensibilidade de 82,2% e F1-score de 83,4% para a identificação de opacidades.

A comparação entre os modelos evidenciou vantagem da EfficientNetB0 em relação à CNN desenvolvida do zero, especialmente na identificação de opacidades, cuja sensibilidade foi de 82,2% na EfficientNetB0 e 57,6% na CNN desenvolvida do zero. Além disso, a CNN treinada do zero apresentou oscilações importantes nas métricas de validação, apesar da melhora progressiva observada durante o treinamento, sugerindo menor capacidade de generalização e possível ocorrência de overfitting.

Apesar dos resultados obtidos, ainda existem limitações para a utilização desses modelos em outros serviços e populações, principalmente pela necessidade de validação

externa e pela influência de diferenças relacionadas às bases de dados, equipamentos e características das imagens. Dessa forma, a ampliação e diversificação do conjunto de dados, o aprimoramento das estratégias de treinamento e a validação em bases independentes são necessários para avaliar de maneira mais robusta o desempenho dos modelos.

Assim, os resultados demonstram que as redes neurais convolucionais, particularmente a EfficientNetB0 associada ao transfer learning, apresentam potencial como ferramentas auxiliares na análise de radiografias de tórax. Entretanto, sua aplicação clínica deve ser considerada como complementar à avaliação profissional, sendo necessários estudos adicionais antes de sua implementação como ferramenta de apoio à triagem e à interpretação de exames radiográficos.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Ludermir TB. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estud Av.* 2021;35:85-94.
2. Santos NM, Campos CFS. Inteligência artificial aplicada à otimização do diagnóstico por imagem. In: Silva R, Neto BRS, editores. *Práticas preventivas e práticas curativas na medicina*. Ponta Grossa: Atena; 2021.
3. Rauber TW. *Redes neurais artificiais*. Vitória: Universidade Federal do Espírito Santo; 2005. p. 39.
4. Pereira NS, et al. Deep learning: conceitos e utilização nas diversas áreas do conhecimento. *Rev Ada Lovelace*. 2018;2:34-49.
5. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-44. doi:10.1038/nature14539.
6. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv [Preprint]*. 2017. doi:10.48550/arXiv.1711.05225.
7. Chartrand G, Cheng PM, Vorontsov E, Drozdal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. *Radiographics*. 2017;37(7):2113-31. doi:10.1148/rg.2017170077.
8. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep convolutional

neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35(5):1285-98. doi:10.1109/TMI.2016.2528162.

9. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. *Adv Neural Inf Process Syst*. 2019;32:3347-57.

10. Salehzadeh A, Calitz AP, Greyling J. A deep learning approach to the diagnosis of multiple abnormalities in chest X-rays. *Bioengineering (Basel)*. 2023;10(11):1317. doi:10.3390/bioengineering10111317.

11. Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge (MA): MIT Press; 2016.

12. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.

13. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell*. 2022;4(3):e210064. doi:10.1148/ryai.210064.

14. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, von Tengg-Kobligk H, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell*. 2020;2(3):e190043. doi:10.1148/ryai.2020190043.