Transfer Learning para Generalização de Redes Neurais Convolucionais em Radiografias de Tórax

Matheus Saick de Martin
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
matheussaick@gmail.com

Pedro Igor Gomes de Morais
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
pedroigorgm@gmail.com

Renzo Henrique Guzzo Leão
Departamento de Informática
Universidade Federal do Espírito Santo
Vitória, Brasil
renzolealguzzo@gmail.com

***Abstract*—A aplicação de redes neurais convolucionais (CNNs) em imagens médicas tem se mostrado promissora, especialmente no diagnóstico de doenças cardiorrespiratórias a partir de radiografias de tórax. No entanto, a capacidade de generalização desses modelos ainda apresenta desafios relevantes quando submetidos a bases de dados distintas. Este trabalho investiga o impacto do *transfer learning* na construção de classificadores, avaliando seu desempenho em dois cenários distintos: (i) treinamento e validação em um conjunto heterogêneo de radiografias; e (ii) avaliação da generalização em uma base de larga escala e composição demográfica distinta. Foram empregados diferentes modelos de CNNs e uma arquitetura baseada em *transformers*, submetidos a *fine-tuning*, validação cruzada e testes estatísticos para verificar diferenças significativas de desempenho. Além disso, análises qualitativas com Grad-CAM foram realizadas para interpretar as regiões de maior influência nas predições. Os resultados indicam que, embora os modelos atinjam métricas elevadas na base de treinamento, há queda expressiva de desempenho ao serem avaliados em um domínio diferente, evidenciando limitações na generalização. A análise sugere que fatores como diferenças etárias, geográficas e de protocolos de aquisição contribuem para esse comportamento, reforçando a necessidade de abordagens mais robustas para a aplicação clínica em larga escala.**

***Index Terms*—redes neurais convolucionais, aprendizado profundo, radiografias de tórax, pneumonia, covid.**

I. Introdução

Na área da saúde, a aplicação de modelos de Inteligência Computacional tem se mostrado promissora, sobretudo em problemas relacionados ao diagnóstico de doenças por meio da análise de radiografias de tórax [1], [2]. Esse tipo de exame desempenha um papel crucial na prática clínica, pois possibilita a detecção de condições de alta incidência, como doenças cardiorrespiratórias e doenças respiratórias crônicas, que demandam diagnósticos rápidos e precisos. A relevância do problema é ainda maior diante do impacto global dessas doenças, que representam significativa carga para os sistemas de saúde [3].

Nesse contexto, técnicas baseadas em redes neurais convolucionais (CNNs) vêm se consolidando como uma das abordagens mais eficazes para a classificação de imagens médicas. Diversos estudos têm demonstrado que o uso de CNNs, aliado a estratégias como *transfer learning*, permite alcançar resultados competitivos em tarefas de diagnóstico a partir de radiografias [4], [5] em diferentes áreas da saúde. Além de contribuir para a precisão do diagnóstico, essas abordagens tem potencial para acelerar a tomada de decisão clínica e apoiar profissionais da saúde em ambientes de alta demanda.

Diante desse cenário, este trabalho tem como objetivo avaliar o uso de modelos pré-treinados quando treinados e testados em bases distintas e como esse processo reflete-se na tarefa de classificação de radiografias de tórax. A proposta é discutir se a abordagem, seguindo as boas práticas de aprendizado de máquina como as apresentadas em [6] e [7], dentre as diversas práticas ressalta-se: a compreensão das bases de dados para contrução de modelos, avaliação de curvas de aprendizado, utilização de técnicas de regularização, entre outras. Bem como a utilização de *transfer learning*, e a avaliação se de fato, o agregado dessas diversas estratégias favorecem ou não a capacidade generalização do modelo, considerando especificamente o cenário onde existe uma clara variação entre as bases de dados analisadas (i.e. a contrução de modelos em uma base e avaliação em outra totalmente distinta). Dessa forma, busca-se compreender melhor a dificuldade dos problemas de classificação de radiografias do tórax e os limites da generalização de modelos treinados nesse domínio.

II. Referencial teórico

A. Definições e conceitos centrais

As radiografias de tórax constituem um dos exames de imagem mais utilizados na prática clínica, sendo amplamente empregadas para avaliação de doenças cardiorrespiratórias devido à acessibilidade, baixo custo e rapidez na obtenção dos resultados [8]. Esse exame desempenha um papel fundamental na triagem inicial, acompanhamento de pacientes e suporte à tomada de decisão médica, especialmente em contextos hospitalares de alta demanda, nos quais a detecção precoce de alterações pulmonares pode impactar diretamente a conduta clínica e os desfechos de saúde [9].

B. Dificuldades no diagnóstico e papel das CNNs

Apesar da relevância, a interpretação de radiografias de tórax apresenta desafios consideráveis. A sobreposição de estruturas anatômicas, a presença de múltiplas doenças concomitantes , erros do observador [10] e as variações morfológicas

decorrentes de sequelas clínicas, dificultam a análise visual precisa e rápida por parte dos especialistas.

Nesse cenário, a aplicação de modelos de redes neurais convolucionais (CNNs) consolidou-se como uma estratégia promissora [1], pois essas arquiteturas são capazes de extrair automaticamente padrões discriminativos a partir de grandes volumes de dados, acelerando do processo de auxílio no diagnóstico e reduzindo a dependência exclusiva da análise manual.

C. Trabalhos relacionados

Diversos estudos têm explorado CNNs na análise de radiografias de tórax, como em [1], que apresentam um levantamento abrangente sobre os avanços recentes, destacando tanto o potencial das redes quanto as limitações em cenários clínicos reais. Importante ressaltar que bases de dados de larga escala, como o *CheXpert* [2] e o *PadChest* [11], foram determinantes para esse avanço ao oferecerem anotações extensas que viabilizaram o treinamento de modelos complexos.

Outras iniciativas, como a de [12], voltada para a COVID-19, expandiram rapidamente a disponibilidade de dados em contextos emergenciais. De forma complementar, em [13] demonstraram a aplicabilidade das CNNs em diferentes patologias, reforçando o caráter generalista da técnica.

O *transfer learning* também se consolidou como uma das abordagens mais eficazes para contornar restrições de dados, já em [14] evidenciaram a eficácia na detecção de COVID-19, enquanto revisões recentes [4], [5] ressaltam a importância de técnicas de *fine-tuning* e reutilização de redes pré-treinadas em diferentes cenários clínicos, principalmente pela escassez de dados. Mais recentemente, trabalhos como o de Mirthipati [15] investigam arquiteturas otimizadas especificamente para doenças torácicas, indicando novos caminhos para o aprimoramento de classificadores.

D. Discussão crítica da literatura

Apesar desses avanços, a literatura revela limitações importantes, visto que em [16] demonstraram que a generalização entre populações distintas ainda é limitada, resultando em quedas de desempenho significativas quando os modelos são treinados e avaliados em grupos diferentes. Do ponto de vista técnico, em [17] ressaltam que variações nos equipamentos e protocolos de exame introduzem ruídos que comprometem a robustez das CNNs, enquanto [18] evidenciaram que a variabilidade entre instituições, práticas hospitalares e configurações de aquisição, podem induzir viés e reduzir a capacidade de generalização.

Além disso, a presença simultânea de múltiplas doenças e de sequelas de condições anteriores pode alterar substancialmente a morfologia das radiografias, dificultando a tarefa de classificação. Esses fatores evidenciam que, embora as CNNs sejam promissoras, a aplicabilidade em cenários clínicos reais ainda enfrenta obstáculos consideráveis que precisam ser superados para ampliar seu uso prático.

III. Metodologia

Tendo em vista os obstáculos supracitados na classificação de radiografias de tórax, neste estudo foram definidos os seguintes passos para a condução dos experimentos:

- Seleção de duas bases de dados distintas. Uma para o treinamento e avaliação do desempenho dos modelos, e outra para testar a capacidade de generalização dos modelos construídos;
- Escolha de arquiteturas de redes neurais amplamente utilizadas e referenciadas em trabalhos correlatos;
- Divisão dos conjunto de dados para o treinamento. Conjuntos de treino, validação e teste, sobre a **Base 1** e também a escolha dos hiperparâmetros;
- Condução de testes estatísticos sobre os modelos obtidos;
- Apresentação de uma análise qualitativa utilizando a técnica GRADCAM [19] sobre a inferência realizada sobre a **Base 2**.

A. Bases de Dados

Duas bases de dados foram selecionadas com o objetivo de inserir heterogeneidade nos cenários de treinamento e avaliação dos modelos construídos, isto é, deliberadamente possuir duas distribuições de dados distintas. Essa escolha, tem por objetivo discutir as barreiras e dificuldades do domínio de aplicação, mesmo quando são utilizadas técnicas e práticas de aprendizado de máquina validadas empiricamente [7] e amplamente referencidas na literatura [6], quando aplicadas na conjuntura proposta.

A primeira base de dados, a **Base 1**, é um apanhado de outros 3 subconjuntos de dados de **imagens de raio-X do pulmão** [20]. Essa base trata-se de um conjunto para classificação multiclasse em três categorias: pneumonia, covid-19 e normal. A contagem total de amostras é 6432 e a distribuição das amostras é a seguinte: a classe pneumonia conta com 3.418 imagens para treino e 855 para teste; a classe normal apresenta 1.266 imagens para treino e 317 para teste; e a classe covid-19 possui 460 imagens para treino e 116 para teste.

A seguir, uma breve apresentação das características gerais de cada subconjunto:

- O primeiro subconjunto é relativo à um projeto da Universidade de Montreal [21]. Esses dados foram utilizados para construção de modelos de classificação de COVID [22], possuindo no total 481 imagens.
- O segundo subconjunto [23] é referente à 5863 imagens de pulmão com pneumonia e pulmão saudável, de crianças de 1 a 5 anos de idade, do hospital *Guangzhou Women and Children's Medical Center*. Neste subconjunto, não foram oferecidos metadados da idade de cada paciente.
- O terceiro subconjunto é referente à imagens de COVID organizadas pelo *Core COVID-Net Team* [24]. Possui no total 48 imagens, sendo que nesse subconjunto são apresentadas a idade para apenas alguns pacientes.

A segunda base de dados, a **Base 2**, é uma amostragem [25] da base **CheXpert** [2], amplamente utilizada em diversos

trabalhos como em [26]–[28]. Essa amostra contém 224.316 imagens, referentes a 65.240 pacientes, cuja distribuição de idade pode ser vista na Figura 1. Sendo notável a variação da faixa etária nesses dados, e o contraste em relação a **Base 1**, isto é, a primeira possui majoritariamente imagens de raio-X de pulmão de crianças (i.e. ainda em estágio de desenvolvimento osseo), já a segundo possui em sua maioria adultos, com uma grande foco para a população acima de 40 anos de idade.

Diferentemente da **Base 1**, que é voltada à classificação multiclasse, a **Base 2** apresenta um cenário de **classificação multi-label**, em que cada imagem pode estar associada a múltiplos rótulos de forma simultânea. Os metadados fornecem 18 características descritivas, entre as quais se destacam 13 diagnósticos de doenças não mutuamente exclusivos — incluindo pneumonia — além da classe *No finding*, que indica ausência de anomalias. Para cada diagnóstico, os rótulos podem assumir quatro valores: **1** (presença da condição), **0** (ausência), **-1** (incerteza na anotação) ou **null** (informação indisponível). Além disso, a base contém metadados adicionais, como a orientação da imagem (frontal ou lateral) e a direção do feixe de raio-X.

Portanto, realizou-se o pré-processamento na **Base 2** a fim de uniformizar os rótulos antes da etapa de avaliação. Nesse processo, os valores **-1** (incerteza) e **None** (ausência de anotação) foram convertidos para **0** (ausência da condição). Essa padronização permitiu a extração e utilização das imagens no conjunto de teste da base, viabilizando a aplicação consistente do classificador e diminuindo possíveis erros da generalização ao limitar apenas em presença ou não do rótulo.

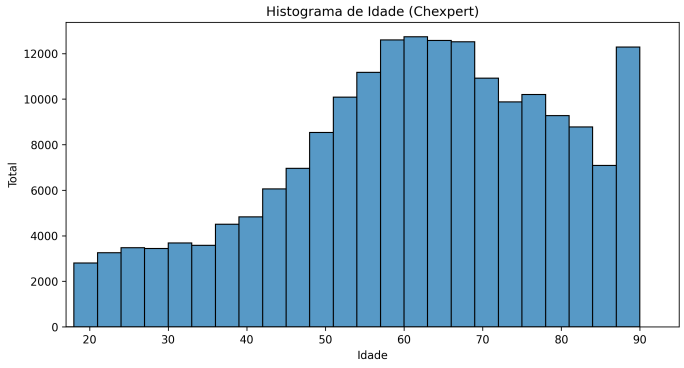

Fig. 1. Distribuição da faixa de idade dos pacientes da **Base 2**

As diferenças entre as bases sobressaem, não apenas relativas ao número de imagens, mas também que incluem variações geográficas e demográficas. No quesito diferenças geográficas, em particular na **Base 1**, o primeiro subconjunto [12] também é uma união de outras bases de dados. Essa união possui dados da *Società Italiana di Radiologia Medica e Interventistica* (uma organização italiana), além de outras fontes de bases abertas disponiveis na Internet. Em contraste, o segundo subcojunto refere-se à dados da cidade de *Guangzhou* na China. Já o terceiro subconjunto, não possui informações expressivas em seus metadados sobre a nacionalidade ou geolocalização do diagnóstico dos pacientes.

Tendo em vista as diferenças notáveis apresentas, foi definido para este trabalho a **Base 1** como o domínio para a construção dos modelos (i.e. o treinamento e avaliação), enquanto a **Base 2** foi utilizada apenas para verificar a capacidade de generalização dos modelos contruídos a partir da primeira base. Essa deliberação, deve-se fato do alto custo computacional de treinar diversos modelos com uma base de dados extensa. Tendo em vista que o volume de dados total utilizado neste trabalho apresenta mais de 200 mil imagens.

B. Seleção de Modelos

A fim de abranger um quantitativo significativo de arquiteturas e realizar um comparativo com grau de expressividade, foram escolhidos os seguintes modelos para *fine-tuning* na **Base 1**: MobileNetV2 [29], MobileNetV3-Large [30], Resnet18 [31], Resnet32 [31], densenet121 [32], densenet161 [32] e ViTb16 [33], esta última baseada na arquitetura *transformer*. A Resnet18 e Resnet32, também foram utilizadas para *feature-extraction*, nomeadas neste trabalho como ”resnet18 frozen” e ”resnet32 frozen”.

C. Divisão dos conjuntos de dados para treinamento

Assim como mencionado anteriormente, a **Base 1** foi definida para a construção dos modelos, sendo que, inicialmente, foram separados os conjuntos de treino e teste. Logo após, foi feito um *k-fold* estratificado de 5 *folds* sobre o conjunto de treino. Vale ressaltar que, durante o treinamento, todas amostras relativas a um paciente ou estavam no conjunto de treino e validação, ou no conjunto de teste, não havendo vazamento de informação.

Ademais, foram definidas estratégias simples de transformações de aumento de dados a fim de auxiliar na robustez dos modelos. Essas transformações são aplicadas exclusivamente no conjunto de treino e as escolhas de cada tipo de transformação aplicada são apresentadas na tabela I.

TABLE I
DATA AUGMENTATION APLICADO NO TREINO DAS REDES CONVOLUCIONAIS

Função	**Parâmetros**
Resize	(224, 224)
RandomHorizontalFlip	p = 0.5
RandomAffine	degrees = 5 translate = (0.05, 0.05)
RandomPerspective	distortion_scale = 0.1
Normalize	mean = [0.5, 0.5, 0.5] std = [0.5, 0.5, 0.5]

D. Testes Estatísticos

Para garantir a robustez da análise comparativa entre os modelos construídos, foi essencial a escolha criteriosa dos testes estatísticos. Optou-se pelo uso dos testes de **Friedman** [34] e **Nemenyi** [35], visto que nenhum dos dois testes exigem que o os dados analisados sigam uma distribuição particular, um aspecto relevante para as bases desbalanceadas escolhidas

para este trabalho. O teste de Friedman visa avaliar a possibilidade de diferença estatística do grupo de classificadores. Já o teste de Nemenyi, busca a verificação da possivel diferença estatística de forma par a par de modelos.

E. Análise qualitativa utilizando GRADCAM

Com o fim avaliar as regiões que mais influenciam na classificação obtida por um modelo em particular, na **Base 2**, foi utilizada a técnica GRADCAM [19]. Essa abordagem foi amplamente utilizada em trabalhos como [36]–[38], a fim de auxiliar na análise dos modelos obtidos e trazer possiveis observações úteis.

IV. Experimentos e Resultados Obtidos

Foi realizado o treinamento dos 9 modelos para cada um dos 5 *folds*, por no máximo 50 épocas, exceto para o modelo "ViTb16", no qual o limite foi definido como 100 épocas, pois, ao longo do treinamento, as curvas de aprendizado dos folds não apresentaram indícios de *overfitting*.

Ademais, foram definidas 5 épocas como *warm-up* e 10 épocas como paciência relativa ao *early stopping* no processo de treinamento das redes. Vale ressaltar que, para os melhores modelos, todos exceto a "ViTb16" encerraram o treinamento antes de completar o número máximo de épocas.

Após a obtenção dos 9 melhores modelos, isto é, o melhor modelo de cada tipo de arquitetura para cada conjunto de 5 *folds*. Foi realizado o teste de Friedman com nível de significância de 5% (α=0.05) aplicado sobre f1-score máximo de cada fold para cada modelo, e foi obtido um p-value que indica possivel diferença estatística no grupo de classificadores. Isto é:

$$p = 0.000043$$

Portanto, tem-se a rejeição da hipótese nula. Com o indicativo de diferença estatística no grupo de classificadores, para cada par, foi realizado o teste de Nemenyi também com α=0.05, como pode ser observado na tabela II, onde é possível observar diferenças estatística entre alguns pares de modelos (p-value's menores que α em negrito). Para complementar a análise, também foi feita uma visualização do gráfico de *Critical Difference* gerado a partir dos valores calculados no teste de Nemenyi, vide Figura 2, com o objetivo de determinar o modelo a ser utilizado para avaliação na **Base 2**. Classificadores que possuem uma distância de *ranking* menor que a *Critical Difference* não apresentam diferença estatística. Esses grupos são representados por linhas mais escuras que cortam as linhas verticais dos classificadores pertencentes ao mesmo grupo. Portanto, a partir do *ranking* do CD, foi selecionado o modelo MobileNetV2 para realização dos experimentos na base CheXpert como melhor modelo (maior *ranking*).

Nota-se que, apesar do desempenho significativo no conjunto de teste da **Base 1**, como pode ser observado pela tabela III, visto que, por exemplo, a média ponderada do *f1-score* se mostra particularmente alta, o modelo não foi capaz de alcançar uma generalização desejada na **Base 2**. Esse comportamento pode ser observado nas Tabelas IV e V, que foram construídas com as predições do modelo escolhido na **Base 2**, tratando cada classe de forma independente. Nesse formato, quando o modelo prediz incorretamente uma classe, o erro é refletido em todas as condições envolvidas. Por exemplo, se para uma amostra rotulada simultaneamente como *Edema* e *Pneumonia* o classificador preve apenas "Covid-19", essa predição incrementa tanto as entradas *(Covid-19, Edema)* quanto *(Covid-19, Pneumonia)* na tabela, que pode ser interpretada como quantidade de chutes para as classes da **Base 2**. Destarte, os valores das tabelas evidenciam a dificuldade do modelo em lidar com a natureza multi-label da **Base 2**, sugerindo um possível *overfitting* na **Base 1**, visto que várias amostras foram dadas como *covid-19*, uma classe com menos instâncias na **Base 1**.

Além da possibilidade de *overfitting* na **Base 1**, podemos formular outras hipóteses sobre a não-generalização do modelo. Assim como foi citado anteriormente, a distribuição de idades da **Base 1** inclui majoritariamente crianças chinesas de 1 a 5 anos. Em contrapartida, a base de dados **CheXpert** possui efetivamente uma população mais velha como observado na Figura 1.

Adicionalmente, foi feita uma análise qualitativa sobre a classificação de amostras pela "mobilenet v2", como demonstrado na Figura 3, cujos rótulos reais são *No finding*, que indica um pulmão que não apresenta as enfermidades apresentadas no conjunto de dados **CheXpert**. Essa análise qualitativa, não evidencia diferenças visuais significativas das doenças, mas sugere a hipótese de que o modelo está prevendo como covid-19 devido à limitação de generalização, isto é, características da nova base de dados são interpretadas como indicativas de covid-19, possivelmente por ser a instância com menos exemplos no treinamento. Isso é enfatizado na classificação inferior à direita da Figura 3, em que o modelo utiliza áreas abaixo da região do tórax para realizar a classificação.

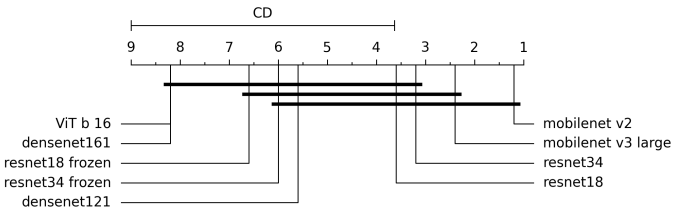

Fig. 2. Gráfico de Critical Difference (CD), uma visualização do teste de Nemenyi sobre a Base 1

V. Conclusões

Neste trabalho foram conduzidos experimentos de comparação de modelos construídos a partir de uma base de dados pequena e particular de um determinado tipo de população. Também foi realizada uma comparação dos modelos contruídos a partir de testes estatísticos e em seguida a avaliação da capacidade de generalização de um dos classificadores desenvolvidos em outra base de dados.

TABLE II
RESULTADOS DO TESTE DE NEMENYI, REPRESENTANDO A SIMILARIDADE ENTRE DIFERENTES ARQUITETURAS DE CNNS E VIT

	resnet18	resnet34	mobilenet v2	mobilenet v3 large	densenet121	densenet161	resnet18 frozen	resnet34 frozen	ViT b 16
resnet18	1.0000	1.0000	0.9035	0.9988	0.9655	0.1639	0.7267	0.9035	0.1639
resnet34	1.0000	1.0000	0.9655	0.9999	0.9035	0.0918	0.5694	0.7960	0.0918
mobilenet v2	0.9035	0.9655	1.0000	0.9988	0.2126	**0.0017**	**0.0476**	0.1238	**0.0017**
mobilenet v3 large	0.9988	0.9999	0.9988	1.0000	0.6501	**0.0230**	0.2701	0.4881	**0.0230**
densenet121	0.9655	0.9035	0.2126	0.6501	1.0000	0.8554	0.9997	1.0000	0.8554
densenet161	0.1639	0.0918	**0.0017**	**0.0230**	0.8554	1.0000	0.9916	0.9400	1.0000
resnet18 frozen	0.7267	0.5694	**0.0476**	0.2701	0.9997	0.9916	1.0000	0.9999	0.9916
resnet34 frozen	0.9035	0.7960	0.1238	0.4881	1.0000	0.9400	0.9999	1.0000	0.9400
ViT b 16	0.1639	0.0918	0.0017	**0.0230**	0.8554	1.0000	0.9916	0.9400	1.0000

TABLE III
TABELA DE MÉTRICAS RELATIVAS AO CONJUNTO DE TESTE NA BASE 1

	covid19	normal	pneumonia	accuracy	macro avg	weighted avg
precision	0.9914	0.9606	0.9734	0.9720	0.9751	0.9719
recall	1.0000	0.9242	0.9859	0.9720	0.9700	0.9720
f1-score	0.9957	0.9421	0.9796	0.9720	0.9724	0.9718
support	116	317	855	0.9720	1288.0000	1288.0000

TABLE IV
PRIMEIRA TABELA DE RESULTADOS DE CLASSIFICAÇÃO NA BASE 2

	No Finding	Enlarged Cardiomediastinum	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation
covid19	16748	9100	23163	93325	6941	49355	12845
normal	133	8	6	84	13	5	9
pneumonia	93	79	216	802	86	315	129

TABLE V
SEGUNDA TABELA DE RESULTADOS DE CLASSIFICAÇÃO NA BASE 2

	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural Other	Fracture	Support Devices
covid19	4635	29508	17409	76092	2479	7366	106133
normal	12	19	39	52	4	12	118
pneumonia	28	193	245	755	22	58	919

Foi observado a dificuldade de generalização em função da diferença entre as distribuições dos dados das duas bases e um possivel *overfitting* sobre a base de treinamento.

Propõe-se como trabalhos futuros a inversão da configuração das bases, ou seja, a definição da primeira base para avaliação da capacidade de generalização dos modelos contruídos a partir da segunda base, bem como a expansão do quantitativo de classificadores analisados. Entre outras possibilidades de continuação do trabalho, vale ressaltar a replicação dos experimentos na mesma configuração prosposta, porém com uma leve extensão da primeira base de dados com um percentual da **Base 2** com o propósito de validar se, com um pequeno conjunto de amostras de outra distribuição, é possível aproveitar o aprendizado de um determinado domínio de aplicação mesmo que as distribuições dos dados sejam distintas. E por fim, sugere-se a incorporação de metadados clínicos juntamente com imagens para o treinamento dos modelos.

REFERENCES

[1] E. Çallı and Colleagues, "Deep learning for chest x-ray analysis: A survey," *Medical Image Analysis*, vol. 72. [Online]. Available: http://dx.doi.org/10.1016/j.media.2021.102125

[2] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: Chest X-Rays," 2019.

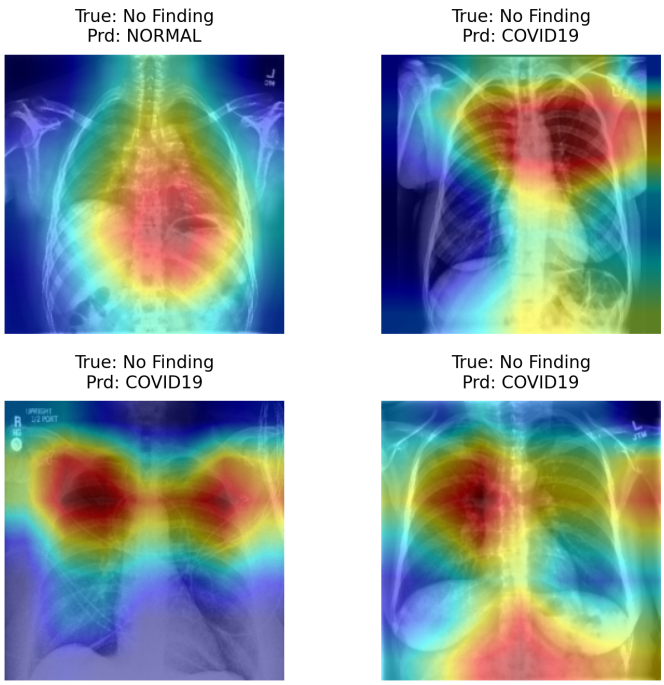

Fig. 3. Visualização da explicabilidade classificação do modelo mobilenet v2 com GradCam

[3] Momtazmanesh and colleagues, "Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the global burden of disease study 2019," *eClinicalMedicine*, vol. 59, p. 101936, May 2023. [Online]. Available: http://dx.doi.org/10.1016/j.eclinm.2023.101936

[4] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med. Imaging*, vol. 22, no. 1, p. 69, Apr. 2022.

[5] P. Kora, C. Ooi, O. Faust, R. U, A. Gudigar, W. Y. Chan, M. Kollati, K. Swaraja, P. Pławiak, and U. Acharya, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics and Biomedical Engineering*, vol. 42, pp. 79–107, 01 2022.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[7] M. J. Zaki and W. Meira, Jr, *Data mining and machine learning*, 2nd ed. Cambridge, England: Cambridge University Press, Jan. 2020.

[8] A. Ganapathy, N. Adhikari, J. Spiegelman, and D. Scales, "Routine chest x-rays in intensive care units: A systemic review and meta-analysis," *Critical care (London, England)*, vol. 16, p. R68, 04 2012.

[9] A. N. Rubinowitz, M. D. Siegel, and I. Tocino, "Thoracic imaging in the icu," *Critical Care Clinics*, vol. 23, no. 3, pp. 539–573, 2007, monitoring in the Intensive Care Unit. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749070407000371

[10] G. F. Abbott, *Challenging Chest Radiograph Interpretation*. Springer Nature Switzerland, 2025, p. 169–187. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-83872-9_14

[11] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical Image Analysis*, vol. 66, p. 101797, Dec. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.media.2020.101797

[12] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[13] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[14] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[15] T. Mirthipati, "Optimizing cnn architectures for advanced thoracic disease classification," 2025. [Online]. Available: https://arxiv.org/abs/2502.10614

[16] M. D. Li, N. T. Arun, M. Aggarwal, S. Gupta, P. Singh, B. P. Little, D. P. Mendoza, G. C. A. Corradi, M. S. Takahashi, S. F. Ferraciolli, M. D. Succi, M. Lang, B. C. Bizzo, I. Dayan, F. C. Kitamura, and J. Kalpathy-Cramer, "Multi-population generalizability of a deep learning-based chest radiograph severity score for COVID-19," *Medicine (Baltimore)*, vol. 101, no. 29, p. e29587, Jul. 2022.

[17] E. a. Fernandez-Miranda, "A retrospective study of deep learning generalization across two centers and multiple models of x-ray devices using COVID-19 chest-x rays," *Sci. Rep.*, vol. 14, no. 1, p. 14657, Jun. 2024.

[18] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, p. e1002683, Nov. 2018. [Online]. Available: http://dx.doi.org/10.1371/journal.pmed.1002683

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," pp. 618–626, 2017.

[20] A. Sani, "Chest x-ray image dataset," https://www.kaggle.com/datasets/alsaniipe/chest-x-ray-image, 2021, accessed: 2025-08-21.

[21] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," *arXiv 2006.11988*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[22] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto, "Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/18/6933

[23] P. Mooney, "Chest x-ray images (pneumonia)," https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia, 2018, accessed: 2025-08-21.

[24] A. Chung and C. C.-N. Team, "Figure1-covid chest x-ray dataset," https://github.com/agchung/Figure1-COVID-chestxray-dataset, 2020, accessed: 2025-08-21.

[25] Ashery, "Chexpert chest x-rays (kaggle sample)," https://www.kaggle.com/datasets/ashery/chexpert/data, 2020, accessed: 2025-08-21.

[26] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221000953

[27] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek, "Comparing different deep learning architectures for classification of chest radiographs," *Scientific Reports*, vol. 10, no. 1, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1038/s41598-020-70479-z

[28] K. Cho, K. Kim, Y. Nam, J. Jeong, J. Kim, C. Choi, S. Lee, J. Lee, S. Woo, G.-S. Hong, J. B. Seo, and N. Kim, "Chess: Chest x-ray pre-trained model via self-supervised contrastive learning," *Journal of Digital Imaging*, vol. 36, 01 2023.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019. [Online]. Available: https://arxiv.org/abs/1801.04381

[30] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," 2019. [Online]. Available: https://arxiv.org/abs/1905.02244

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: https://arxiv.org/abs/1608.06993

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[34] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, p. 675–701, Dec. 1937. [Online]. Available: http://dx.doi.org/10.1080/01621459.1937.10503522

[35] P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. thesis, Princeton University, 1963.

[36] L. S. Chow, G. S. Tang, M. I. Solihin, N. M. Gowdh, N. Ramli, and K. Rahmat, "Quantitative and qualitative analysis of 18 deep convolutional neural network (cnn) models with transfer learning to diagnose covid-19 on chest x-ray (cxr) images," *SN Computer Science*, vol. 4, no. 2, Jan. 2023. [Online]. Available: http://dx.doi.org/10.1007/s42979-022-01545-8

[37] A. Alqutayfi, W. Almattar, S. Al-Azani, F. A. Khan, A. A. Qahtani, S. Alageel, and M. Alzahrani, "Explainable disease classification: Exploring grad-cam analysis of cnns and vits," *Journal of Advances in Information Technology*, vol. 16, no. 2, p. 264–273, 2025. [Online]. Available: http://dx.doi.org/10.12720/jait.16.2.264-273

[38] A.-K. Balve and P. Hendrix, "Interpretable breast cancer classification using cnns on mammographic images," 2024. [Online]. Available: https://arxiv.org/abs/2408.13154