Cross-dataset application for fair evaluation of COVID-19 X-ray models

Gabriel de S. G. Pedroso[1], **Thiago M. Ventura**[1], **Allan G. de Oliveira**[1]

[1]Instituto de Computação – Universidade Federal do Mato Grosso (UFMT)
Cuiabá – MT – Brasil

gabriel.de.s.g.pedroso@gmail.com, {thiago,allan}@ic.ufmt.br

Abstract. *Various efforts were undertaken for the recognition of COVID-19 through X-ray images. The studies achieved good performance in recognizing these images. However, the models are tailored to the datasets they are trained on, which does not imply the same performance outside the training context. Therefore, this study applied a fair way for evaluating models across diverse scenarios. The results demonstrated that the models managed to differentiate among different datasets from various sources, thus, it was determined that the conducted studies were adapted to the context of the acquired dataset.*
Keywords*: x-ray. evaluation. model. coronavirus. protocol*

Resumo. *Diversos trabalhos foram realizados para o reconhecimento de COVID-19 por meio de imagens de raio-X. Os trabalhos obtinham bom desempenho no reconhecimento das imagens, no entanto, os modelos estão alinhado aos conjuntos de dados utilizados, o que não implica o mesmo desempenho fora do contexto de treino. Deste modo, este trabalho aplicou uma forma justa para avaliar modelos em diferentes cenários. Os resultados demonstraram que os modelos conseguiram distinguir entre diferentes conjuntos de dados de origem diferente, assim, foi determinado que os trabalhos realizados estiveram adaptados ao contexto do conjunto de dados obtido.*
Palavras-chave*: raio-x. avaliação. modelos. coronavírus. protocolo*

1. Introdução

Com o surgimento da COVID-19 e seu avanço pelo mundo, diversos trabalhos com Inteligência Artificial (IA), em específico, as Redes Neurais Convolucionais (*Convolutional Neural Networks - CNN*), começaram a ser realizados para auxiliar no combate da doença [Shorten et al. 2021, Altan and Karasu 2020].

Dentro do propósito de reconhecimento de doenças por imagens de raio-X, surgem problemas relacionados ao viés destes classificadores para ser possível o uso de ambientes reais. Isso é retratado em [Maguolo and Nanni 2021], no qual critica os resultados dos classificadores e que novas formas de avaliação da capacidade de generalização dos modelos treinados devem ser criadas. Ainda com esta preocupação, em [Guarrasi et al. 2022] é proposto um método *Ensemble* com o uso de diversos modelos de CNNs para fortificar os classificadores, demonstrando as melhorias sobre dados de outros *datasets*.

Esse trabalho propõe a realização de validação-cruzada entre diferentes abordagens para definição dos conjuntos de dados de imagens de raio-X, mantendo como base

a separação original e comparando os resultados alcançados para validar os esforços e vieses para reconhecer COVID-19. Com isso, espera-se a criação de uma forma na qual os resultados dos classificadores representem melhor as condições em ambientes reais de utilização.

2. Metodologia

2.1. Conjunto de dados

Foram selecionadas duas bases para este trabalho. A primeira delas foi estabelecida por [Mooney 2017] contém 5.848 imagens com diferente resoluções em pixel em formato jpeg de crianças saudáveis ou com pneumonia e está disponível em `https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia`. A outra base, com imagens de COVID-19, foi proposta por [Cohen et al. 2020] e possui 866 imagens com diferentes resoluções em pixel em formato jpeg e png, disponível em `https://github.com/ieee8023/covid-chestxray-dataset`. Esta base também contém metadados para cada imagem, como o doutor que fez *upload* da imagem, sexo da pessoa, localização de onde o raio-X foi tirado, entre outros. Todas as imagens foram utilizadas em escala de cinza, redimensionadas para 224x224 pixels e normalizadas entre 0 e 1. A Tabela 1 contém os dados resumidos das bases.

Base de dados	Pneumonia	Saudável	COVID-19
Cohen	147	0	504
Kag	4273	1583	0

Tabela 1. Número de imagens de cada classe nas bases de dados

2.2. Definição de modelos e escolha de *Transfer Learning*

Foram definidos dois modelos para a captura de diferentes aspectos das imagens e um modelo para *Transfer Learning*. O primeiro modelo foi definido para realizar a captura de regiões mais largas das imagens de raio-X, enquanto o segundo foi definido para extrair diferentes características (de alto nível a baixo nível) da região torácica. As ideias de captura dos modelos baseiam-se na tentativa, respectivamente, de manter e não manter a dimensão das imagens após uma camada convolucional, de modo que no primeiro modelo sempre toda a imagem é processada, e no segundo algumas regiões, a fim de ir especificando ao longo das camadas as características ideias para reconhecimento.

Todos os filtros foram utilizados com dimensão 3x3. Seguem 2 blocos convolucionais com função de ativação *ReLU*, cada um composto por 1 camada convolucional, o primeiro com 16 filtros e o 2º com 32 filtros, acompanhada por outra de *Max Pooling*. Logo após, 20% dos pesos aleatoriamente são zerados, antes de prosseguir para o último bloco convolucional com 64 filtros. Por fim, há a camada totalmente conectada de 128 neurônios com função de ativação *ReLU* e a última camada para classificação de 7 classes codificadas com *one-hot-encoding*: situação normal, pneumonia bacteriana e pneumonia viral para a base *Kag*, pneumonias virais, por fungos, bacterianas e a COVID-19 da base *Cohen*. No restante do texto, ambos os modelos serão mencionados como *Larger Regions* (LR) e *Low-High Level* (LHL), respectivamente.

Para o modelo com *Transfer Learning* [Diment and Virtanen 2017] foi utilizado o *Resnet 50* [He et al. 2016], uma vez que bons resultados foram obtidos em diferentes segmentos, inclusive com imagens de raio-X, como pode ser visto em [Guarrasi et al. 2022]. Todas as camadas foram congeladas, adaptando os pesos somente para a camada de classificação deste trabalho.

2.3. Forma de avaliação dos modelos

Os dados foram separados da seguinte forma: 70% para treino, 20% para validação e 10% para teste. Com os modelos treinados, foram calculados a precisão e *recall*. A fase de treinamento ocorreu durante 15 épocas.

Após uma primeira avaliação, foram realizadas mais três iterações visando: aplicação de validação cruzada com os dados de treinamento e validação para cada modelo da etapa anterior; e avaliação da distinção de subclasses das possíveis visualizações do raio-X; separação dos 10% dos dados para teste enquanto o mesmo restante da 2º iteração foi utilizado para a validação cruzada de 5 *folds*.

3. Resultados e Discussões

A Tabela 2 demonstra menor capacidade de reconhecimento da base *Cohen* a qual possui menor quantidade de dados. Houve algumas imagens as quais foram identificadas incorretamente à base *Kag*, visto que o *recall* de *Kag* foi de 100% no teste, de modo que a qualidade de distinção da maior parte das imagens com a base *Cohen* foi inferior no teste. Os resultados para os modelos criados apresentaram-se iguais, enquanto o *Resnet 50* demonstrou, a princípio, efetivamente ter aprendido o padrão dos dados, o que tornou possível reconhecer a qual base pertence cada imagem. Já a Tabela 3 demonstra a melhoria de precisão para a base *Cohen* em detrimento da capacidade de detecção durante a validação. Apesar desta melhoria, a precisão teve queda acentuada com o conjunto de teste para a base *Kag*, além da diminuição do *recall* para a base *Cohen*, o que demonstra *overfitting* do modelo.

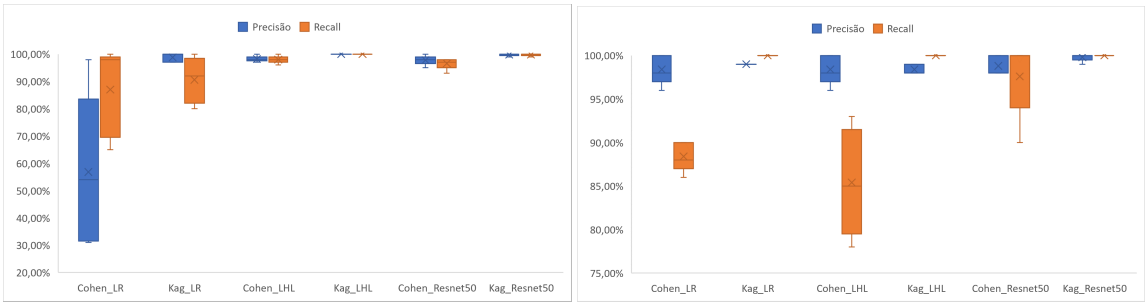
Base	Conjunto	Precisão			*Recall*		
		LHL	LR	*Resnet 50*	LHL	LR	*Resnet 50*
Cohen	Validação	91%	90%	94%	95%	95%	99%
Cohen	Teste	100%	100%	100%	97%	97%	100%
Kag	Validação	99%	99%	100%	98%	98%	99%
Kag	Teste	89%	89%	100%	100%	100%	100%

Tabela 2. Métricas da 1º iteração para o reconhecimento de cada base de dados com os modelos escolhidos

Base	Conjunto	Precisão			*Recall*		
		LHL	LR	*Resnet 50*	LHL	LR	*Resnet 50*
Cohen	Validação	92%	94%	96%	88%	89%	96%
Cohen	Teste	100%	100%	100%	84%	92%	95%
Kag	Validação	98%	98%	99%	99%	99%	99%
Kag	Teste	62%	76%	84%	100%	100%	100%

Tabela 3. Métricas da 3º iteração para o reconhecimento de cada base de dados com os modelos escolhidos

Por meio das mudanças de conjunto de dados sobre o modelo, foram observados os impactos sobre os mesmos dados. Os resultados para a 2º iteração podem ser visualizados na Figuras 1a demonstrando a precisão e *recall*, com os modelos como sufixo para cada *plot*. Há grande variabilidade na precisão para a base *Cohen*, com um pouco menos no *recall*, ainda assim, é possível verificar que houve maior sensibilidade para o modelo LR com a base, enquanto o modelo demonstrou maior uniformidade para a base *Kag*. O modelo LHL apesar de ter demonstrado também menor sensibilidade a mudanças para a base *Kag*, obteve resultados superiores ao outro modelo, visto que a distribuição percentual nem ao menos foi inferior a 95%. Em comparação ao modelo LHL, o *Resnet 50* demonstrou menor qualidade e capacidade de detecção, com a distribuição das métricas bem semelhantes ao modelo, com superioridade ao modelo LR.



(a) **(b)**

Figura 1. Box plot da distribuição da precisão e *recall* para (a) os modelos da 2ª iteração e (b) as bases *Cohen* e *Kag* da 4ª iteração

Para a validação cruzada com a utilização de metadados, é possível verificar na Figura 1b para o modelo LHL e para o *Resnet 50* a obtenção de menor desempenho para ambas as bases de dados. No entanto, para o modelo LR, foi constatada menor variabilidade a diferentes subconjuntos das bases de dados, com melhoria de desempenho do modelo.

Levando em consideração todos esses resultados, é possível observar diferentes comportamentos para reconhecimento de imagens de raio-X de acordo com a base utilizada. Com a definição da forma de avaliação, não somente as métricas foram utilizadas para verificar o comportamento diferente sobre os dados, mas também subconjuntos diferentes e uso de metadados mostraram impactos no modelo. A forma de avaliação estabelecida demonstrou o quão direcionado um modelo está em relação a própria base de dados a qual foi utilizada e que o seu estabelecimento pode auxiliar a enxergar possíveis limitações, a fim dez melhorar o desempenho de um classificador.

4. Considerações Finais

Os resultados deste trabalho demonstraram a aderência dos modelos criados para conseguir reconhecer as bases de dados escolhidas. O passo a passo pode ser verificado no repositório do *github* em `https://github.com/SousaPedroso/ERI-MT-2023`.Eles também demonstraram o benefício de aplicar a forma de avaliação, visto que foi possível associar por cada classificador a base alvo. Isso é um alerta para a necessidade de validação de modelos e a utilização de conjuntos de dados de diferente origens para a possibilidade de uso em produção.

Para trabalhos futuros, podem ser incorporadas mais bases para avaliação, como a *NIH* [Wang et al. 2017] ou a apresentada por [Irvin et al. 2019]. Também é possível realizar modificações nas imagens de cada base a fim de verificar quais possíveis regiões de uma imagem de raio-X tiveram mais impacto para o reconhecimento das bases ou quais atenções devem ser tomadas de modo a padronizar diferentes conjuntos de dados para que não haja o reconhecimento da base de origem.

Referências

Altan, A. and Karasu, S. (2020). Recognition of covid-19 disease from x-ray images by hybrid model consisting of 2d curvelet transform, chaotic salp swarm algorithm and deep learning technique. *Chaos, Solitons & Fractals*, 140:110071.

Cohen, J. P., Morrison, P., and Dao, L. (2020). Covid-19 image data collection. arXiv preprint arXiv:2003.1159.

Diment, A. and Virtanen, T. (2017). Transfer learning of weakly labelled audio. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 6–10. IEEE.

Guarrasi, V., D'Amico, N. C., Sicilia, R., Cordelli, E., and Soda, P. (2022). Pareto optimization of deep networks for covid-19 diagnosis from chest x-rays. *Pattern Recognition*, 121:108242.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597.

Maguolo, G. and Nanni, L. (2021). A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 76:1–7.

Mooney, P. (2017). Chest x-ray images (pneumonia). `https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia`. Acesso em: 10 ago. 2023.

Shorten, C., Khoshgoftaar, T. M., and Furht, B. (2021). Deep learning applications for covid-19. *Journal of big Data*, 8(1):1–54.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.