Classificação de Distúrbios Pulmonares em Radiografias de Tórax Usando Redes Convolucionais

Alysson Machado, Leo Araújo e Luciana Veloso

***Resumo*— A radiologia depende da extração de informações em imagens, sendo uma área de aplicação natural para aprendizado profundo, cujos modelos se destacam sobretudo em tarefas de Visão Computacional. Destarte, investiga-se a utilização de redes convolucionais para auxiliar radiologistas na classificação dos distúrbios pulmonares: opacidade, lesão, edema, consolidação, atelectasia, pneumotórax e efusão. São exploradas arquiteturas renomadas de redes convolucionais e comitês de classificação, produzindo resultados satisfatórios na classificação multirrótulo de radiografias, com AUC médio de 83,49%. Em análise subsequente, é feita uma avaliação sobre a relação dos mapas de ativação dos modelos em exames laudados por radiologistas.**

***Palavras-Chave*— Visão Computacional, Redes Neurais Convolucionais, Aprendizado de Máquina, Distúrbios Pulmonares, Classificação Multirrótulo.**

***Abstract*— Radiology depends on extracting information from images, being a natural application area for deep learning, whose models stand out above all in Computer Vision tasks. Thus, we investigate the use of convolutional networks to assist radiologists in the classification of pulmonary disorders: opacity, lesion, edema, consolidation, atelectasis, pneumothorax and effusion. Renowned architectures of convolutional networks and classification committees are explored, producing satisfactory results in the multi-label classification of radiographs, with an average AUC of 83,49%. In a subsequent analysis, an evaluation is made of the relationship of the activation maps of the models in exams reported by radiologists.**

***Keywords*— Computer Vision, Convolutional Neural Networks, Machine Learning, Pulmonary Disorders, Multi-label Classification.**

I. Introdução

Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNN), técnica pertencente ao domínio da aprendizagem profunda, possui a capacidade de aprender a classificar imagens a partir do treinamento com outras grandes quantidades de imagens rotuladas, sendo uma das suas principais características a extração de informações úteis a partir desses tipos de dados [1]. O diagrama esquemático deste tipo de rede é ilustrado na Figura 1.

Atualmente existem inúmeras arquiteturas renomadas baseadas em CNN, como *DenseNet [2], Inception [3], Xception [4], InceptionResnet [5], ResNet [6], VGG [7] e MobileNet [8]*, que permitem o aumento da performance desses modelos, treinados com quantidade modesta de dados, em aplicações distintas das originais, utilizando uma técnica chamada Transferência de Aprendizado (*Transfer Learning*) [9].

Alysson Machado, Departamento de Engenharia Elétrica, UFCG, Campina Grande-PB, e-mail: alysson.barbosa@ee.ufcg.edu.br; Leo Araújo, Departamento de Engenharia Elétrica, UFCG, Campina Grande-PB, e-mail: leo.araujo@ee.ufcg.edu.br; Luciana Veloso, Departamento de Engenharia Elétrica, UFCG, Campina Grande-PB, e-mail: luciana.veloso@dee.ufcg.edu.br.

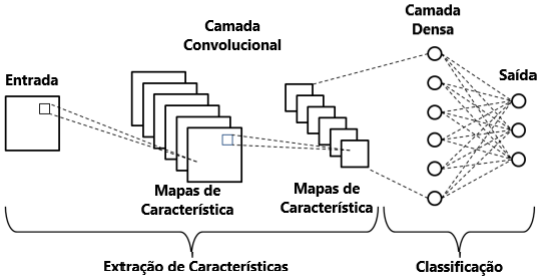


Fig. 1: Diagrama Esquemático de uma Rede Neural Convolucional. Retirado de [10].

Dentre as áreas da medicina que podem se beneficiar com as CNNs, a classificação de distúrbios pulmonares em radiografias mostrou-se ser proeminente [11], principalmente ao considerar o agrave das enfermidades na área da Pneumologia no atual cenário mundial. Segundo o programa de pesquisa regional e global *Burden of Disease*, em seu estudo publicado em 2015 [12], levando em consideração várias métricas de saúde em tendências globais, distúrbios respiratórios são considerados um dos problemas mais recorrentes em todo mundo, abrangendo uma vasta gama de doenças responsáveis por comprometer o sistema respiratório do indivíduo. Além disso, constituem 5 das 30 causas mais comuns em relação à mortalidade hospitalar. Outro problema recorrente, principalmente em países menos desenvolvidos, é que, embora haja a presença de máquinas de raio-x, a ausência ou insuficiência de profissionais com experiência radiológica para avaliar com precisão as imagens médicas, inviabiliza o uso em potencial dessas máquinas. Em contraste a isso, nos países mais desenvolvidos, a radiografia de raio-x é utilizada para fazer a triagem de pacientes e a determinação dos níveis de cuidados requeridos a eles, exigindo um gasto considerável na análise manual dessas imagens. Outrossim, com o advento da epidemia de SARS-CoV-2, houve um aumento na demanda de soluções tecnológicas para possibilitar a atuação remota de uma parcela dos radiologistas mais experientes, tendo em vista as diretrizes de isolamento social [13].

Dentro desse contexto, este projeto investiga a utilização das CNNs para auxiliar os profissionais da medicina na classifica-

ção de doenças pulmonares utilizando imagens radiológicas, as quais constituem um problema de classificação multirrótulo, em que uma mesma imagem pode pertencer a mais de uma classe. Em específico, diferentes modelos de CNNs foram analisados para determinação de suas contribuições, utilizando um mesmo conjunto de métricas e uma mesma base de dados. Além disso, foi montado comitês de classificação (*Ensemble Learning*) [14] a partir dos modelos produzidos para analisar os potenciais ganhos de se utilizar os diversos classificadores simultaneamente, juntamente com a construção de mapas de ativação de classe para verificar a qualidade dos modelos.

II. Metodologia

Para avaliar os modelos treinados com a estratégia de transferência de aprendizado, a base de dados selecionada foi particionada em três conjuntos: treino, validação e teste. A performance dos modelos com relação à classificação das patologias representadas na base de dados foi avaliada a partir de métricas avaliativas como Tempo de Inferência, Curva ROC (*Operating Characteristic Curve*) e AUC (*Area Under the ROC Curve*), cujo valor médio será indicado por $\overline{AUC}$. Além disso, foram analisados os mapas de ativação, obtidos por meio do algoritmo *Grad CAM* [15], em comparação com as anotações realizadas por radiologistas em exames de uma base de dados independente.

A. Base de Dados

CheXpert é um conjunto de imagens públicas [16] de radiografias de tórax feita para competições na área de interpretação automatizada das radiografias torácicas, que apresentam laudos multirrótulo. As imagens consistem em 224.316 radiografias de tórax provenientes de 65.240 pacientes. Tais dados foram coletados de exames radiográficos de tórax do *Stanford Hospital*, realizados entre outubro de 2002 e julho de 2017, tanto em centros de internação quanto em ambulatórios.

As imagens utilizadas para treinamento, validação e teste dos modelos foram extraídas a partir da base de dados *CheXpert*. Algumas amostras de imagens foram removidas por não se adequarem aos experimentos realizados. Como o foco do trabalho é a classificação de distúrbios pulmonares em radiografias frontais, amostras com projeção lateral e amostras associadas a distúrbios que não afetam os pulmões foram removidas. Além disso, foram retiradas radiografias referentes a pacientes com menos de 10 anos ou mais de 80 anos, bem como imagens da classe "pneumonia", cujo número de amostras é muito pequeno se comparado às demais. Outrossim, foi realizado um controle sobre a quantidade de radiografias de cada classe, visando a obtenção de partições balanceadas entre as sete classes de distúrbios pulmonares selecionadas: opacidade, lesão, edema, consolidação, atelectasia, pneumotórax e efusão, com um saldo final de 25.164 imagens selecionadas para alimentar os modelos (rotulações de incerteza sobre a presença de um determinado distúrbio não foram consideradas). Por fim, 80,0% dessas imagens foram reservadas para Treinamento (20.131 imagens), 10,0% para Validação (2.517 imagens) e 10,0% para Teste (2.517 imagens), todas amostragens com aproximadamente o mesmo grau de estratificação. A quantidade de rótulos por instância nas imagens selecionadas é apresentada na Figura 2.

B. Aumento de Dados

Antes de encaminhar as imagens para a entrada das redes, o aumento de dados (*Data Augmentation*) foi aplicado no subconjunto de treinamento. Os seguintes procedimentos foram considerados: mudança de escala, deformações suave, aproximações suaves e deslocamentos para esquerda e direita.

Tais processos simulam um banco de dados maior e mais diverso, mitigando os efeitos do sobreajuste (*overfitting*) [17]. Exemplos das imagens de treinamento após o aumento de dados são apresentados na Figura 3. As estratégias de aumento de dados utilizadas aplicam transformações de baixo impacto para preservar os padrões contidos nos dados, evitando descaracterizar os exemplos utilizados [18].

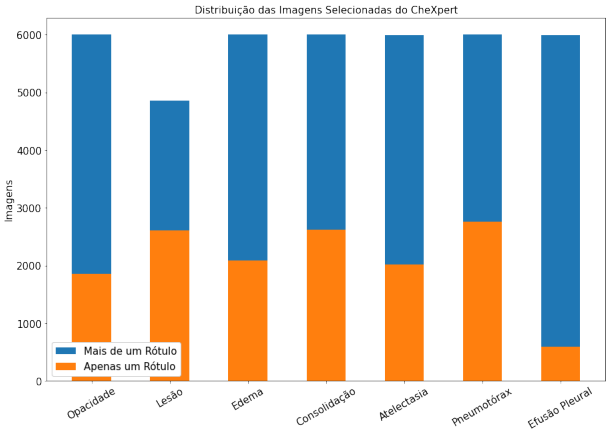


Fig. 2: Distribuição das Imagens retiradas da Base de Dados *CheXpert* laudadas por radiologistas. **Mais de um Rótulo**: imagens com mais de um laudo associado. **Apenas um Rótulo**: imagens com apenas um rótulo associado.

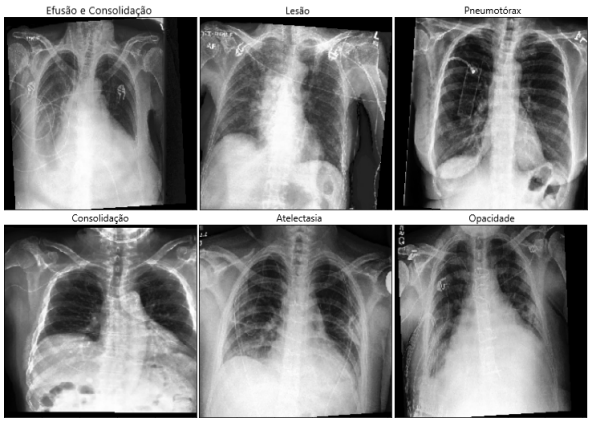


Fig. 3: Exemplos de radiografias com a aplicação do aumento de dados (*Data Augmentation*).

C. Treinamento

O treinamento das redes [19] foi realizado utilizando a API *Keras* em conjunto com o *Framework TensorFlow*, utilizando

a linguagem de programação Python. O ambiente de desenvolvimento foi o *Google Colaboratory PRO*, com disponibilidade de GPUs da *Google Compute Engine*.

Todos os modelos foram construídos tendo como base arquiteturas renomadas na literatura do *Deep Learning*. São elas: *DenseNet [2], Inception [3], Xception [4], InceptionResNet [5], ResNet [6], VGG [7] e MobileNet [8]*. As imagens foram configuradas com as dimensões 256x256 ao longo dos três canais RGB. Em relação a inicialização dos pesos da rede, foi utilizada a estratégia de Transferência de Aprendizado (*Transfer Learning*) [9] a partir do treinamento com conjuntos de imagens da *ImageNet*. Foram aproveitadas as camadas convolucionais dos modelos, que foram sucedidas por uma camada de *Global Average Pooling* e uma camada totalmente conectada com ativação sigmóide para realização da classificação.

Em relação à compilação da rede, a função de custo foi definida como sendo a *Binary Crossentropy*, considerando a natureza binária de cada saída, que classifica a imagem quanto à presença ou ausência de uma dada enfermidade. A métrica de validação realizada ao longo das 10 épocas utilizadas no treinamento de cada modelo foi a métrica $\overline{AUC}$ com base nas imagens de validação. O *Adam* foi escolhido como otimizador, fazendo uso de uma taxa de aprendizado com um valor inicial de 1.10^{-4}, escolhido através de testes empíricos, sendo ajustado automaticamente em caso de platô com base nos valores de $\overline{AUC}$ computados com os dados de validação, multiplicando o seu valor por uma taxa de 0.6 a cada ajuste se necessário.

D. Comitês de Classificação

Visando obter um melhor desempenho preditivo, combinou-se os modelos produzidos em dois comitês de classificação. A primeira estratégia utiliza o método de média da predição de cada um dos modelos sem ponderação, enquanto a segunda estratégia utiliza o método de média da predição de cada um dos modelos ponderada pela métrica $\overline{AUC}$ de cada classificador.

E. Avaliação Quantitativa

Avaliações quantitativas foram realizadas nos modelos a partir das respectivas métricas:

(I) **Curva ROC**: representação gráfica da performance de um modelo binário à medida que o limiar de discriminação dos classificadores varia [20].
(II) **AUC**: quantifica a área abaixo da curva ROC [20].
(III) **Tempo de Inferência**: informa o tempo necessário que o modelo demanda para classificar uma imagem.

F. Avaliação Qualitativa

A realização da avaliação qualitativa dos modelos foi feita através da utilização do algoritmo *Grad Cam*. Esse algoritmo explora informações contidas no gradiente integrado das CNNs para produzir um mapa de ativações, possibilitando a atribuição da previsão realizada pelo modelo a elementos da sua entrada. Dessa forma, é possível gerar uma máscara para cada uma das imagens em análise, visando indicar as áreas mais influentes para a predição dos modelos [15]. Um modelo bem treinado para classificação de distúrbios pulmonares possui tais áreas influentes concentradas na região do pulmão, indicando o local em que existe uma anormalidade associada. As imagens utilizadas nesta análise foram extraídas de um banco de dados independente hospedado no site *Radiopaedia* [21], que contém uma série de radiografias anotadas por radiologistas.

III. Resultados e Discussões

Os valores das avaliações quantitativas e qualitativas foram calculados a partir dos dados de teste, imagens que permaneceram separadas da etapa de treinamento e validação.

A. Avaliação Quantitativa

Analisando os gráficos da curva ROC, ilustrados na Figura 4 e os dados disponibilizados de $\overline{AUC}$ na Tabela I, é possível observar que as redes treinadas possuem performance semelhante, ainda que uns se destaquem mais do que outros, como a *InceptionResnetV2*. Levando em conta as diferenças arquiteturais dos modelos produzidos, segue-se que a forma como cada um deles extrai características a partir dos dados de entrada difere, o que é confirmado pela variação no valor de $\overline{AUC}$ desses modelos, os quais foram treinados, validados e testados através do mesmo conjunto de imagens. Dada essa distinção, foi possível obter melhores predições utilizando estratégias de comitês de classificação. Essa relação é evidenciada pelo aumento de AUC em todas as classes na Figura 4(h) com relação aos modelos individuais da Figura 4(a-g).

A Tabela I reúne as métricas $\overline{AUC}$ dos modelos e o tempo de inferência utilizado na análise de uma única imagem. Uma análise interessante é que a arquitetura *DenseNet121* demanda um tempo de inferência de 69,46 milissegundos apresentando um $\overline{AUC}$ de 76,78%, enquanto a arquitetura *InceptionResNetV2*, mais otimizada [5], necessita de um tempo de inferência de 64,50 milissegundos e possui um valor de $\overline{AUC}$ de 80,65%. O tempo de inferência está associado ao grau de otimização da arquitetura, a qual as estratégias utilizadas na conexão das camadas profundas de convolução impactam diretamente na melhora dessa métrica. Além disso, a *InceptionResNetV2* apresenta um número maior de camadas profundas (164 camadas) em relação às camadas da *DenseNet121* (121 camadas). Em análise subsequente, a quantidade de camadas profundas das CNN não garantem necessariamente que a rede consiga extrair as melhores informações da base de dados, como é possível observar na comparação da taxa de $\overline{AUC}$ com a quantidade de camadas das arquiteturas *Xception* e *InceptionV3*, e também a *VGG16* e *ResNet101V2*.

A Figura 5 mostra os valores de AUC das classes para cada um dos classificadores. Esses resultados indicam que os modelos possuem resultados melhores quando utilizados em conjunto, assim como permitem que as predições sejam realizadas por modelos com arquiteturas de CNNs distintas, fornecendo resultados mais consistentes em cada uma das classes, como é possível observar na taxa de $\overline{AUC}$ da média simples de 83.47% e da média ponderada de 83.49%. Concomitantemente a isso, o tempo de inferência dessas estratégias aumentou de forma abrupta. Todavia, comparado ao tempo

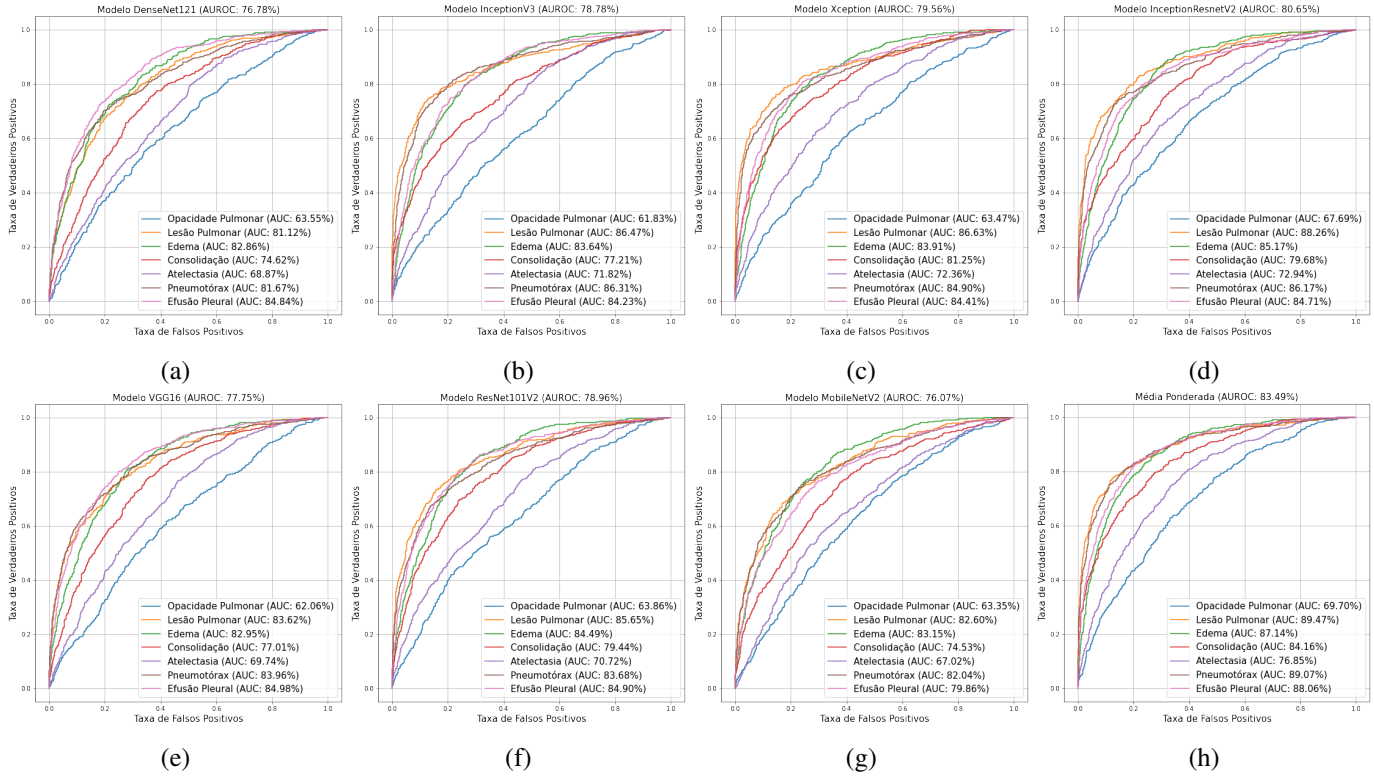


Fig. 4: Curvas ROC com os respectivos valores de AUC por classe e $\overline{AUC}$ calculados no conjunto de teste para cada arquitetura Figuras (a-g) e para o comitê de classificação por média ponderada Figuras (h).

de análise manual de um profissional radiologista, o tempo de inferência para as classificações automatizadas mostram-se superiores [22]. Em uma análise individual dos classificadores, a arquitetura *InceptionResnetV2* apresentou melhores resultados, com uma taxa de $\overline{AUC}$ de 80,65%, apresentando resultados satisfatórios de AUC para cada uma das classes, com um destaque para opacidade e lesão pulmonar.

TABELA I: Análise Quantitativa dos modelos gerados.

Modelos	$\overline{AUC}$ (%)	Inferência (ms)	Camadas
DenseNet121	76,78	69,46	121
InceptionV3	78,78	52,06	48
Xception	79,56	44,84	36
InceptionResnetV2	80,65	64,50	164
ResNet101V2	77,75	54,55	34
VGG16	78,96	41.65	16
MobileNetV2	76,07	47,67	53
Média Simples	83.47	301,96	--
Média Ponderada	83.49	336,72	--

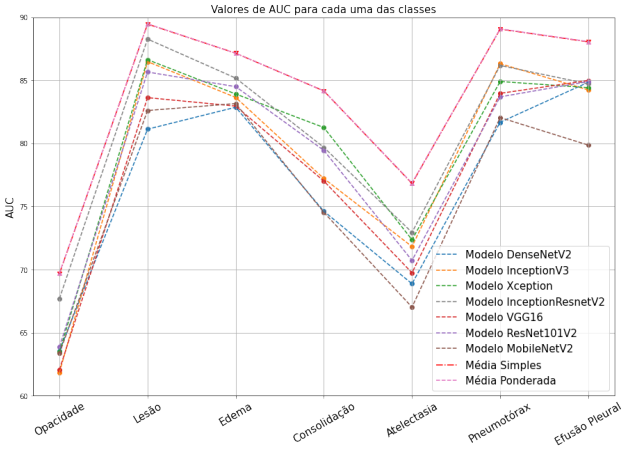


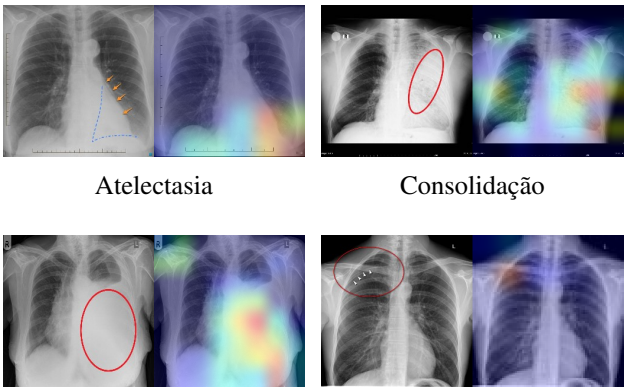
Fig. 5: Valores de AUC das classes por classificador.

Destaca-se que o objetivo de ferramentas desenvolvidas neste trabalho não tem o objetivo de automatizar completamente a classificação de distúrbios pulmonares, mas sim permitir que ferramentas de diagnóstico auxiliado por computador sejam desenvolvidas, com o objetivo de melhorar a produtividade do profissional e a triagem adequada do paciente.

B. Avaliação Qualitativa

No experimento qualitativo, os modelos foram testados com radiografias laudadas por radiologistas utilizando o algoritmo *Grad Cam* [15], de modo a comparar se as áreas de ativação das classes nas imagens correspondem às mesmas áreas destacadas por radiologistas. Entretanto, é perceptível que os modelos ainda aprendem padrões a partir de regiões desconexas da área pulmonar. Tais inconsistências derivam da grande heterogeneidade existente nas radiografias, a qual se mostram desafiadoras para os modelos aprenderem todas as nuances envolvidas na análise de distúrbios pulmonares. Assim como podem advir das condições em que as radiografias foram coletadas e também da forma como o algoritmo *Grad Cam* foi proposto (pesquisas realizadas posteriormente ao estágio de desenvolvimento do projeto analisa a existência de algoritmos

mais precisos [23]). Outrossim, é possível que os profissionais da radiologia extraiam novos conhecimentos a partir dos resultados observados por essa categoria de algoritmos, pois existem características presentes nas radiografias não discerníveis por inspeção visual conhecidas como informações radiômicas [24], às quais modelos baseados em CNN conseguem identificar com mais facilidade. Os resultados deste algoritmo são ilustrados na Figura 6.



Atelectasia Consolidação

Efusão pulmonar Pneumotórax

Fig. 6: Análise qualitativa com o algoritmo *Grad Cam* em radiografias com anotações de radiologistas.

IV. Considerações Finais

Essa pesquisa demonstra estratégias úteis para a utilização das Redes Neurais Convolucionais na classificação de distúrbios pulmonares em radiografias torácicas. Apesar do tamanho moderado dos conjuntos de dados utilizados, técnicas como transferência de aprendizado, aumento de dados e comitês de classificação, produziram resultados satisfatórios para classificação de doenças pulmonares de forma automatizada.

Os modelos concebidos podem ser refinados para a efetiva aplicação na área médica como ferramentas complementares. Nesse sentido, a realização de treinamento nessas redes utilizando conjunto de dados mais robustos e consistentes tornaria-os capazes de lidar com a grande heterogeneidade existente nas imagens médicas, permitindo a produção de novas tecnologias para otimizar, de forma complementar, o trabalho dos profissionais da radiologia, assim como garantir aos pacientes uma triagem mais adequada. Além disso, pesquisas futuras podem analisar a interpretabilidade e reusabilidade de algoritmos inteligentes na área da pneumologia, assim como técnicas de extração das informações úteis nos conjuntos de imagens públicas de forma otimizada utilizando mais de uma base de dados. Dessa forma, é possível observar que as redes neurais convolucionais colaboram para o reconhecimento automático de distúrbios pulmonares e corroboram para a expansão da pesquisa no campo da visão computacional utilizando imagens médicas.

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através do programa PIBIC/CNPq-UFCG, e também pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), por meio do processo 88881.507204/2020-01, aprovado no Edital CAPES no âmbito do Edital Emergencial Nº 12/2020.

Referências

[1] ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. *Understanding of a convolutional neural network*. In: IEEE. 2017 International Conference on Engineering and Technology (ICET). [S.l.], 2017. p. 1–6.

[2] HUANG, Gao; LIU, Zhuang; VAN DER MAATEN, Laurens. Kilian Q, Weinberger. *Densely Connected Convolutional Networks*, 2018.

[3] SZEGEDY, Christian et al. *Going deeper with convolutions*. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1-9.

[4] CHOLLET, François. Xception: *Deep learning with depthwise separable convolutions*. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 1251-1258.

[5] SZEGEDY, Christian et al. *Inception-v4, inception-resnet and the impact of residual connections on learning*. In: Thirty-first AAAI conference on artificial intelligence. 2017.

[6] HE, Kaiming et al. *Deep residual learning for image recognition*. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-778.

[7] SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[8] HOWARD, Andrew G. et al. *Mobilenets: Efficient convolutional neural networks for mobile vision applications*. arXiv preprint arXiv:1704.04861, 2017.

[9] PAN, S. J.; YANG, Q. *A survey on transfer learning*. IEEE Transactions on knowledge and data engineering, IEEE, v. 22, n. 10, p. 1345–1359, 2009.

[10] PHUNG, Van Hiep et al. *A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets*. Applied Sciences, v. 9, n. 21, p. 4500, 2019.

[11] DUNNMON, J. A. et al. *Assessment of convolutional neural networks for automated classification of chest radiographs*. Radiology, Radiological Society of North America, v. 290, n. 2, p. 537–544, 2019.

[12] WANG, H. et al. *Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015*. The lancet, Elsevier, v. 388, n. 10053, p.1459–1544, 2016.

[13] ALVIN, Matthew D. et al. *The impact of COVID-19 on radiology trainees*. 2020.

[14] SAGI, Omer; ROKACH, Lior. *Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 8, n. 4, p. e1249, 2018.

[15] SELVARAJU, Ramprasaath R. et al. *Grad-CAM: Why did you say that?*. arXiv preprint arXiv:1611.07450, 2016.

[16] IRVIN, Jeremy et al. *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019. p. 590-597.

[17] MCBEE, M. P. et al. *Deep learning in radiology*. Academic radiology, Elsevier, v. 25, n. 11, p. 1472–1480,2018.

[18] EVANGELISTA, Lucas Gabriel Coimbra; GUEDES, Elloá B. *Ensembles of Convolutional Neural Networks on Computer-Aided Pulmonary Tuberculosis Detection*. IEEE Latin America Transactions, v. 17, n. 12, p. 1954-1963, 2019.

[19] MACHADO, Alysson. *CNN lung Diseases*. GitHub. Disponível em: <https://github.com/Alyssonmach/cnn-lung-diseases>. Acesso em: 08 de jun. de 2021.

[20] FAWCETT, Tom. An introduction to ROC analysis. Pattern recognition letters, v. 27, n. 8, p. 861-874, 2006.

[21] RADIOPAEDIA. *Radiopaedia Organization*. Disponível em: <https://radiopaedia.org/>. Acesso em: 08 de jun. de 2021.

[22] JAISWAL, Amit Kumar et al. *Identifying pneumonia in chest X-rays: a deep learning approach. Measurement*, v. 145, p. 511-518, 2019.

[23] ALGHAMDI, Hanan et al. *Deep learning approaches for detecting COVID-19 from chest X-ray images: A survey*. IEEE Access, 2021.

[24] THRALL, J. H. et al. *Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success*. Journal of the American College of Radiology, Elsevier, v. 15, n. 3, p. 504–508, 2018.