

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

MATHEUS PAES DE SOUZA

EXPLICABILIDADE DE MODELOS DE VISÃO COMPUTACIONAL APLICADOS
À DETECÇÃO DE COVID-19 EM RADIOGRAFIAS

Rio de Janeiro - Brasil
2020

MATHEUS PAES DE SOUZA

**EXPLICABILIDADE DE MODELOS DE VISÃO COMPUTACIONAL APLICADOS
À DETECÇÃO DE COVID-19 EM RADIOGRAFIAS**

Trabalho de conclusão de curso apresentado à Escola de Matemática Aplicada, como parte dos requisitos para a obtenção do título de Bacharel em Matemática Aplicada.

Orientador: Jorge Poco

Rio de Janeiro - Brasil

2020

MATHEUS PAES DE SOUZA

**EXPLICABILIDADE DE MODELOS DE VISÃO COMPUTACIONAL APLICADOS
À DETECÇÃO DE COVID-19 EM RADIOGRAFIAS**

Trabalho de Conclusão apresentado à Escola de Matemática Aplicada como requisito para a
obtenção do grau de bacharel em Matemática Aplicada

Aprovado em 07 de Dezembro de 2020



Jorge Poco

Escola de Matemática Aplicada



Moacyr A. H. Barbosa da Silva

Escola de Matemática Aplicada



Elizabeth Wegner Karas

Universidade Federal do Paraná

Resumo

No final de 2019, Covid-19, uma nova doença respiratória de grande capacidade contagiosa e letal, se espalhou pelo planeta, alcançando status de pandemia. Diante da falta de testes, foram propostos modelos de visão computacional para a detecção da doença a partir de imagens de radiografias, utilizando algoritmos de inteligência artificial treinados em datasets curados para esse objetivo. No entanto, não é incomum a introdução de vieses, presentes de maneira desconhecida nos datasets, no modelo de predição. Neste trabalho, esse fenômeno é investigado, treinando modelos para a detecção de Covid-19 em radiografias e aplicando técnicas recentes de interpretabilidade para modelos de visão computacional (Grad-CAM, Grad-CAM++) a modelos propostos para a detecção da doença.

Palavras-chave: Covid-19, aprendizado de máquina, visão computacional, aprendizado de máquina interpretável, classificação.

Abstract

At the end of 2019, Covid-19, a novel respiratory disease with high contagious and lethal power, spread throughout the planet, reaching pandemic status. Due to the shortage of tests, computer vision models were proposed for the detection of the disease from X-rays, utilizing artificial intelligence algorithms trained on datasets created for this purpose. However, it is not uncommon for biases, unknowingly present in the datasets, to be introduced in the prediction model. In this work, this phenomenon is investigated, training models for the detection of Covid-19 from X-rays and applying recent interpretability techniques for computer vision models (Grad-CAM, Grad-CAM++) to proposed models for the detection of the disease.

Keywords: Covid-19, machine learning, computer vision, interpretable machine learning, classification.

Sumário

| | | |
|----------|--------------------------------------------------------|-----------|
| 1 | Introdução | 7 |
| 1.1 | Motivação | 8 |
| 1.1.1 | Os problemas das redes convolucionais | 8 |
| 1.1.2 | O problema dos datasets | 8 |
| 1.1.3 | Explicabilidade de CNNs e suspeitas iniciais | 9 |
| 1.2 | Objetivos | 9 |
| 1.3 | Considerações finais | 10 |
| 2 | Background em redes neurais convolucionais | 11 |
| 2.1 | Redes neurais artificiais | 11 |
| 2.2 | A inovação das redes convolucionais | 12 |
| 2.3 | Estrutura das redes convolucionais | 13 |
| 2.4 | O desenvolvimento das redes convolucionais | 15 |
| 2.5 | Aplicações das redes convolucionais | 18 |
| 2.6 | Técnicas de explicabilidade | 18 |
| 2.6.1 | CAM | 19 |
| 2.6.2 | Grad-CAM | 19 |
| 2.6.3 | Grad-CAM++ | 20 |
| 2.6.4 | LIME | 21 |
| 2.7 | Considerações finais | 21 |
| 3 | Trabalhos relacionados | 22 |
| 3.1 | Radiografias torácicas | 22 |
| 3.2 | Tomografias computadorizadas e ultrassom | 24 |
| 3.3 | Considerações finais | 24 |
| 4 | Metodologia | 25 |
| 4.1 | Construção do dataset | 25 |
| 4.1.1 | Datasets disponíveis | 25 |
| 4.1.2 | <i>Data Augmentation</i> | 26 |
| 4.1.3 | Dataset com segmentação dos pulmões | 27 |
| 4.2 | Treinamento dos modelos classificadores | 28 |
| 4.3 | Métodos para interpretabilidade | 28 |
| 4.4 | Considerações finais | 29 |

| | | |
|----------|-------------------------------------------------------------------|-----------|
| 5 | Resultados | 30 |
| 5.1 | Avaliação de métricas de classificação | 30 |
| 5.1.1 | Acurácia | 30 |
| 5.1.2 | Precisão | 31 |
| 5.1.3 | Revocação (<i>recall</i>) | 32 |
| 5.1.4 | Valor F | 33 |
| 5.1.5 | Matriz de Confusão | 33 |
| 5.2 | Análise de interpretabilidade | 36 |
| 5.2.1 | Avaliação empírica das explicações | 36 |
| 5.2.2 | Comparação das explicações com observações radiológicas | 39 |
| 5.3 | Considerações finais | 40 |
| 6 | Conclusão | 41 |
| 7 | Referências | 42 |

1 Introdução

Em dezembro de 2019, um surto de pneumonia causado por uma doença infecciosa até então desconhecida teve origem na cidade de Wuhan, província de Hubei, na China [1]. A doença, posteriormente nomeada Covid-19, se espalhou rapidamente pelo planeta ao longo dos próximos meses, alcançando o status de pandemia [2]. Até 15 de novembro de 2020, foram reportados mais de 5,8 milhões de casos e 160 mil mortes causadas por Covid-19 no Brasil. Em todo o mundo, mais de 53,7 milhões de casos e 1,3 milhão de mortes foram reportados [3].

Covid-19 é uma doença respiratória causada pelo vírus SARS-CoV-2, da família dos coronavírus. Pacientes tem sintomas como febre e tosse seca, assim como, embora em menor quantidade, dificuldades respiratórias, dores musculares, diarreia e dor de cabeça. Em casos graves, pacientes podem apresentar pneumonia e síndrome respiratória aguda grave [4], necessitando de internação hospitalar e da utilização de ventiladores mecânicos, que auxiliam na respiração. Indivíduos infectados por SARS-CoV-2 podem também não apresentar sintomas até o fim da infecção.

Embora agora estejam disponíveis diversos testes para Covid-19, com variados níveis de confiabilidade, no início da pandemia os únicos métodos disponíveis e confiáveis para o diagnóstico de casos de Covid-19 eram testes baseados em **transcrição reversa seguida de reação em cadeia da polimerase** (RT-PCR) para detectar o RNA do SARS-CoV-2. A **transcrição reversa** é um processo que sintetiza uma cadeia de DNA a partir de RNA, enquanto a **reação em cadeia da polimerase** é um processo que replica um segmento específico de DNA milhares ou milhões de vezes, resultado em uma quantidade detectável deles. Esse processo, no entanto, é demorado, e pode levar algumas horas, além de requerer equipamento especializado. Além disso, a súbita e enorme demanda por esses testes levou ao acúmulo de testes para serem feitos, e à escassez de matéria-prima para os testes.

Devido à alta infecciosidade do SARS-CoV-2 e alta taxa de mortalidade da Covid-19, o diagnóstico das pessoas infectadas com o vírus é de grande importância, para que se dê início a procedimentos de quarentena, observação e auxílio médico, se necessário. Alguns primeiros estudos sugeriram o uso de radiografias ou tomografias computadorizadas (CT) dos pulmões como uma alternativa aos testes RT-PCR em falta [5]. Diante disso, logo foram também propostos diversos modelos de visão computacional utilizando redes neurais convolucionais para analisar as radiografias e detectar a presença de uma infecção pelo vírus. A literatura já descreve abordagens semelhantes, para a detecção de pneumonia em radiografias torácicas ou exames de CT [6].

1.1 Motivação

A classe de modelos utilizada pelas abordagens e a escassez de dados causada pela novidade da Covid-19 apresentam alguns desafios para o sucesso da detecção automática de Covid-19 em radiografias. Além disso, modelos desenvolvidos para essa tarefa devem ser interpretáveis para que seus resultados possam ser confiáveis.

1.1.1 Os problemas das redes convolucionais

Todas as abordagens propostas utilizam redes neurais convolucionais (CNNs). CNNs são modelos extremamente complexos, com milhões de parâmetros, e são notórias pela alta capacidade de *overfitting*, pela demanda por enormes quantidades de dados durante o treinamento e pela dificuldade em oferecer interpretabilidade dos modelos. Esses são pontos importantes, pois a novidade da doença significava que havia poucos dados disponíveis para o treinamento de um modelo para a detecção de Covid-19, e um modelo treinado deve ser cuidadosamente inspecionado e estudado antes de ser implantado em qualquer ambiente médico, para que se tenha compreensão de seu funcionamento interno.

Modelos como regressões lineares, regressões logísticas, árvores de decisões e *K-Nearest Neighbors* são inerentemente interpretáveis, ainda que simples. Por outro lado, modelos baseados em redes neurais artificiais e, em particular, redes convolucionais, não têm explicação natural para seus resultados. Tais modelos são compostos por uma enorme quantidade de parâmetros que se relacionam de maneiras variadas e são permeados de não-linearidades.

1.1.2 O problema dos datasets

Devido à novidade da Covid-19, não existia nenhum dataset robusto e estabelecido de radiografias especificamente criado para o treinamento de modelos de detecção de Covid-19. Pesquisadores passaram a realizar *scrapping* de páginas na internet e de artigos científicos que publicavam radiografias já anotadas como positivas para Covid-19 para construir seus próprios datasets. Para obter radiografias de pacientes saudáveis ou com outros tipos de pneumonia, eles se tornaram para datasets já existentes, como *NIH Chest X-ray Dataset*, o dataset disponibilizado por Mendeley e CheXpert.

Com a combinação de diversos datasets distintos, cada um com possíveis particularidades, como postura dos pacientes na radiografia e notas para identificação, pode ser possível aprender o dataset de origem. Além disso, a presença de setas demarcando regiões de interesse nas radiografias, anotações de diagnóstico, dispositivos médicos em pacientes internados e escala para medição são possíveis indicadores do diagnóstico de uma radiografia, o que

é chamado de *data leakage*.

Dada a natureza propensa a *overfitting* das redes convolucionais, é perfeitamente possível que os modelos estejam aprendendo a procurar por esses artefatos, em vez de aprender os sinais característicos nos pulmões que indicariam o diagnóstico. Claramente, um modelo no primeiro cenário não seria em nada útil em um contexto real, onde não se sabe previamente o diagnóstico de uma radiografia.

1.1.3 Explicabilidade de CNNs e suspeitas iniciais

Apesar de redes convolucionais serem chamadas de modelos *caixas pretas* no contexto de interpretabilidade, isso não quer dizer que não há nenhuma maneira de investigar os mecanismos internos desses modelos. Pode-se empregar técnicas de interpretabilidade que são *model-agnostic* (ou que não dependem do modelo), como, por exemplo, LIME, que utiliza a imagem de entrada para identificar regiões importantes para o modelo. Outras técnicas foram desenvolvidas especialmente para lidar com redes convolucionais, como CAM, Grad-CAM, Grad-CAM++ e *Guided Backpropagation*.

Em uma postagem em um blog [7], Abdul Khader Jilani, da Data Robots, realizou análises em dois datasets, construídos com radiografias de pacientes com Covid-19 e radiografias de pacientes sem Covid-19. Jilani encontrou evidências de que seus modelos estavam aprendendo que marcações das letras L ou R na radiografia eram indicativas de um diagnóstico negativo, assim como a presença do pescoço, mandíbula e ombros. Além disso, no dataset de radiografias de Mendley, um dos datasets comumente utilizados em tarefas de detecção de pneumonia através de radiografias, e incorporado em um popular dataset para Covid-19, consiste de 5856 radiografias que foram coletadas de crianças. Isso significa que sua utilização juntamente com outros datasets, que não contém radiografias pediátricas, possivelmente introduziria um viés no classificador, dada a grande diferença morfológica entre imagens típicas de cada dataset.

1.2 Objetivos

Como notado, redes convolucionais, que são complexas e propensas a *overfitting*, foram utilizadas na confecção de modelos para o diagnóstico de Covid-19 em radiografias do tórax, e Jilani identificou a possibilidade de *leakage* nos datasets. Tais modelos devem ser interpretados, para que se possa ter confiança em seus resultados.

Neste trabalho, pretendemos analisar o comportamento dos modelos propostos, de maneira a confirmar a qualidade de suas detecções ou identificar possíveis problemas com elas. Para isso, tentaremos reproduzir alguns dos trabalhos relacionados, aplicar técnicas de in-

interpretabilidade nos modelos propostos, treinar modelos aplicando técnicas para melhorar interpretabilidade e prover uma análise dos resultados obtidos.

1.3 Considerações finais

Esta seção descreveu a pandemia de Covid-19, métodos de diagnóstico da doença e o papel de radiografias torácicas nesse processo. Também introduziu o uso de redes convolucionais para detectar Covid-19 em radiografias, junto com as dificuldades e desvantagens de seu uso e a necessidade de interpretabilidade dos modelos. Finalmente, ficaram descritos os objetivos do trabalho. A próxima seção define redes neurais e convolucionais e introduz conceitos utilizados ao longo do trabalho.

2 *Background* em redes neurais convolucionais

Atualmente, redes neurais convolucionais são extensivamente utilizadas em aplicações de visão computacional. Seu espantoso desempenho combinado com avanços em técnicas computacionais permitiu um vertiginoso desenvolvimento na área nos últimos anos. Redes convolucionais formam um classe particular de redes neurais artificiais. A seguir, introduzimos os conceitos de redes neurais artificiais e redes convolucionais, os avanços ocorridos nos últimos anos e esforços para desenvolver métodos para explicar seus resultados.

2.1 Redes neurais artificiais

A unidade básica das redes neurais artificiais é um neurônio artificial. Estes, semelhantemente aos neurônios biológicos, são construções que recebem ativações (valores de saída) de outros neurônios e computam uma ativação em resposta, que é encaminhada para outros neurônios. O valor de um neurônio depende de um valor de entrada x fornecido como dado de entrada, ou de valores de entrada x_i que vêm da saída de outros neurônios conectados a esse. Esses valores são combinados linearmente com pesos w_i e somados a uma constante b , chamada de viés, e o resultado é então introduzido em uma função não linear, chamada de função de ativação. Essa função é responsável por transformar a saída de um neurônio e introduzir não-linearidade ao modelo. Funções frequentemente utilizadas na ativação são, por exemplo, as sigmóides, para mapear a saída linear do neurônio ao intervalo $[0, 1]$; a função ReLU, definida por $ReLU(x) = \max(x, 0)$, que suprime valores negativos, de maneira remanescente ao funcionamento dos neurônios no cérebro; a função softmax para classificação multi-classe. Formalmente, dadas as entradas x_i , com $i \in I = \{1, \dots, n\}$, os pesos w_i de cada conexão, o viés b e uma função de ativação σ , o valor de ativação a de um neurônio é $a = \sigma\left(b + \sum_{i \in I} w_i x_i\right)$.

Redes neurais artificiais são modelos constituídos de neurônios artificiais, dispondo-os em camadas e estabelecendo relações (ou conexões) inter-camada e intra-camada. Geralmente, as camadas são dispostas sequencialmente, e diz-se que uma rede é mais “profunda” quanto mais camadas possui. Os neurônios da primeira camada recebem os dados de entrada, e a última fornece os valores de saída da rede. As camadas intermediárias são também referidas como camadas ocultas (do inglês *hidden layers*). Uma rede neural na qual todas as conexões são entre neurônios de camadas diferentes em direção ao final da rede é caracterizada como pró-alimentada ou *feed-forward*, ao passo que se a ela possuir conexões formando um ciclo a rede é então chamada de rede neural recorrente (RNN, em inglês).

Uma rede neural artificial pró-alimentada é completamente conectada, ou *fully connected*, quando suas conexões se dão somente entre neurônios de camadas adjacentes e todos os

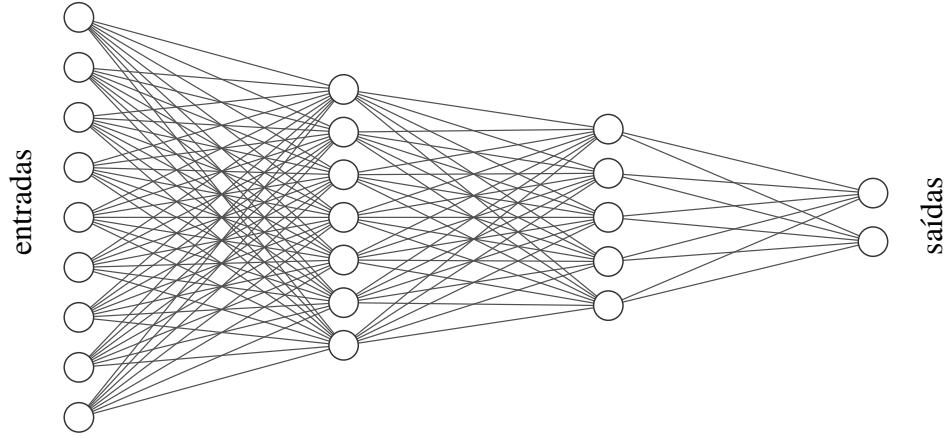


Figura 1: Diagrama de uma rede neural completamente conectada¹

neurônios de uma camada estão conectados com todos os neurônios da seguinte. A Figura 1 exibe, como exemplo, o diagrama de uma rede completamente conectada. Esse tipo de rede neural é utilizado como parte das redes convolucionais. A seguir, formalizamos a modelagem de redes completamente conectadas.

Denotando por n_L o número de neurônios da L -ésima camada; por $b_j^{(L)}$ e $a_j^{(L)}$ o viés e a ativação, respectivamente, do j -ésimo neurônio da L -ésima camada; por $w_{ij}^{(L)}$ o peso da conexão entre o i -ésimo neurônio da $(L-1)$ -ésima camada e o j -ésimo neurônio da L -ésima camada; e por σ_L a função de ativação utilizada na L -ésima camada, descrevemos o valor da ativação $a_j^{(L)}$:

$$a_j^{(L)} = \sigma_L \left(b_j^{(L)} + \sum_{i=1}^{n_{L-1}} w_{ij}^{(L)} a_i^{(L-1)} \right).$$

Os pesos $w_{ij}^{(L)}$ e $b_j^{(L)}$ são parâmetros da rede, requerendo otimização. A abordagem dominante para treinamento das redes neurais artificiais é a aplicação de um algoritmo de otimização baseado no método do gradiente descendente em uma função de custo, utilizando gradientes obtidos por *back-propagation*. *Back-propagation* é uma técnica que utiliza a Regra da Cadeia para calcular, de maneira eficiente, os gradientes da função de custo com relação aos parâmetros da rede neural.

2.2 A inovação das redes convolucionais

Antes da revolução provocada pelas redes convolucionais, a abordagem tradicional para a classificação de imagens era um sistema composto por duas partes [9, 10].

¹Imagem gerada com NN-SVG [8]

A primeira parte era responsável por realizar extração de *features* das imagens, isto é, a transformação da imagem entrada (um vetor de alta dimensão) em uma representação por um vetor de baixa dimensão. Vetores com menores dimensões são mais facilmente manipuláveis e representações mais simples podem ser feitas relativamente invariantes às diversas maneiras como um padrão pode se apresentar. Os componentes para a extração de *features* eram feitos à mão, requerendo conhecimento prévio sobre o domínio a serem aplicados e eram específicos para cada tarefa. O desempenho do modelo dependia em grande parte do desenvolvimento de boas *features*, o que era uma tarefa imensamente difícil.

A segunda parte consistia em um classificador treinável, que não precisava ser específico para cada tarefa. Esse classificador processava o vetor de *features* e produzia uma classificação.

Redes neurais convolucionais introduzem uma modificação na primeira parte. CNNs também consistem em dois módulos: um extrator de *features* e um classificador treinável. No entanto, diferente dos sistemas anteriores, o treinamento não é restrito apenas ao classificador. Em vez disso, a extração de *features* é feita por uma rede neural, que também passa por treinamento, aprendendo a partir dos dados de treino quais são as melhores *features* a serem extraídas. Isso elimina a necessidade de que *features* sejam confeccionadas à mão para cada aplicação e de conhecimento específico no domínio. A desvantagem é a necessidade de enormes quantidades de dados de treinamento para que a rede seja então capaz de desenvolver conhecimento verdadeiro sobre o domínio. Redes convolucionais se tornaram tão grandes e complexas que elas são capazes de simplesmente “decorar” pequenos datasets de treino, alcançando grande precisão nos dados de treino mas não sendo capazes de generalizar esse desempenho para dados fora da amostra de treino.

2.3 Estrutura das redes convolucionais

Uma rede neural convolucional é composta por um módulo responsável por extrair *features*, utilizando redes neurais com camadas *convolucionais*, seguido por um módulo para classificar as *features*, geralmente uma rede neural multi-camada completamente conectada.

Especificamente no caso de redes convolucionais voltadas para processamento de imagens, os neurônios de cada camada convolucional podem ser imaginados como arranjos em uma matriz de três dimensões $m \times n \times c$, composta por c mapas de *features*, de dimensões m e n , da camada anterior. Cada um dos c mapas de *features* também é chamado de um canal. A Figura 2 ilustra uma camada convolucional com 3 canais (i.e., $c = 3$). A primeira camada possui um canal para cada canal das imagens de entrada; por exemplo: 1 canal para preto e branco, 3 para imagens coloridas com RGB ou ainda mais para imagens de sensoriamento

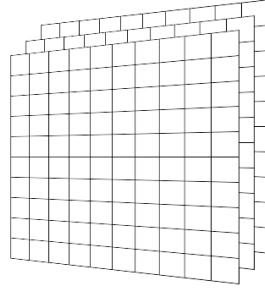


Figura 2: Representação de uma camada convolucional com 3 canais

remoto. Nas camadas seguintes, o número de canais pode ser alterado.

Diferentemente das redes neurais completamente conectadas, as camadas convolucionais das CNNs não se conectam por completo. Em vez disso, cada neurônio de uma camada interna (ou convolucional) tem conexões locais, tendo como entradas as saídas de neurônios localizados em uma pequena vizinhança na camada anterior. Tal vizinhança tem a forma de um bloco $k_1 \times k_2 \times c$, que “desliza” pela camada de acordo com o neurônio à qual se conecta, como ilustrado na Figura 3. Os pesos dessas conexões são reutilizados em cada bloco para neurônios de um mesmo canal e camada. Cada uma das c matrizes $k_1 \times k_2 \times c$ de pesos é chamada de kernel ou filtro, e a operação de aplicar um kernel varrendo a camada, que resulta no mapa de *features*, é chamada convolução. Se o caminho dos filtros não for contínuo, pulando um determinado número de neurônios a cada passo, o resultado será um mapa de *features* de menor dimensão que o original. A saída da convolução é então passada como entrada para uma função de ativação.

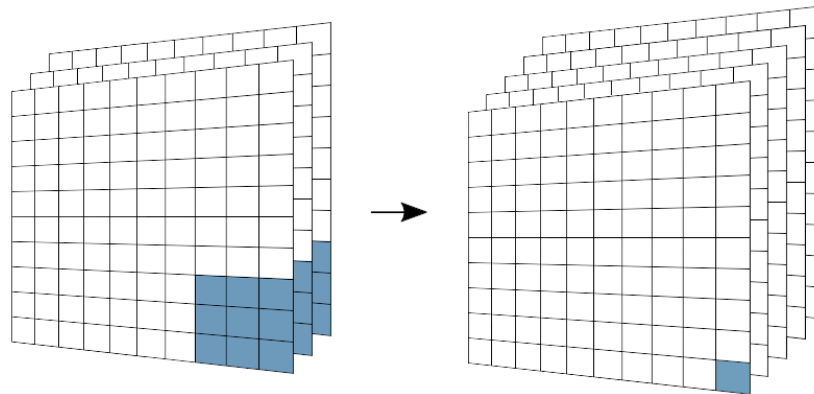


Figura 3: Conexões em uma camada convolucional

Essa estrutura é inspirada na hipótese de que um kernel útil para a extração de *features* em uma parte da imagem provavelmente também é útil em outras partes da imagem, e de que a localização exata das *features* não é importante, mas sim sua localização em relação a *outras features*. Isso decorre do fato de que padrões podem se apresentar levemente deslocados em

diferentes imagens, como em imagens de dígitos numéricos, ou em posições completamente arbitrárias da imagem, como na tarefa de detecção de objetos.

Entre as camadas convolucionais, frequentemente coloca-se uma camada de *pooling*, para reduzir a resolução dos mapas de *features* da camada anterior. Em cada canal, essa redução é feita através da substituição de uma vizinhança de neurônios por uma computação com seus valores de ativação, de maneira semelhante às camadas convolucionais, com um passo maior que 1. Computações comumente utilizadas são tomar a média ou o máximo entre os valores, conhecidas por *average pooling* e *max pooling*, respectivamente. A substituição das informações refinadas por um resumo das ativações vizinhas também trabalha sob o princípio da não importância das localizações exatas das ativações.

2.4 O desenvolvimento das redes convolucionais

Redes neurais convolucionais foram propostas pela primeira vez em 1989, como parte de um modelo para reconhecimento de imagens com desempenho de ponta, por LeCun et al. [11, 12]. Eles propuseram uma pequena rede convolucional com cerca de 2600 parâmetros livres, chamada LeNet-1, para classificar dígitos escritos à mão representados em imagens de 16x16 pixels, com quase nenhum processamento prévio das imagens. O treinamento foi realizado utilizando *back-propagation*, ainda quando esse algoritmo estava ganhando popularidade como ferramenta para o treinamento para redes neurais, e foram capazes de alcançar apenas 1,7% de erro no conjunto de teste. Esse modelo foi aplicado com sucesso para ler códigos ZIP (como são chamados os códigos postais nos Estados Unidos) escritos à mão. Em 1998, LeCun et al. apresentaram LeNet-4 e LeNet-5, novas redes com aperfeiçoamentos à arquitetura anterior e um maior tamanho [9]. Experimentos com a LeNet-1 os convenceram de que redes maiores conseguiriam melhor proveito do grande tamanho de seu dataset de dígitos. A arquitetura LeNet-5 conseguiu 0,95% de erro de teste com o dataset sem modificações. Utilizando um dataset aumentado de 60.000 imagens para 600.000 imagens obtido aplicando combinações aleatórias de distorções como translação, escalonamento e cisalhamento às imagens originais, eles diminuíram o erro de teste para 0,8%. O modelo foi implantado em diversos bancos nos Estados Unidos para ler quantias escritas em cheques, tanto à mão quanto por máquinas. Em [9], LeCun et al. também publicaram MNIST, um grande dataset com imagens de dígitos escritos à mão e suas classes correspondentes.

Por vários anos, dificuldades enfrentadas com redes neurais diminuíram o ritmo de pesquisa nessa área. Pesquisadores amplamente acreditavam que a tarefa de aprender extratores de *features* úteis em múltiplos estágios com pouco conhecimento prévio do domínio não era factível porque a otimização pelo algoritmo do gradiente descendente simples ficaria presa

em mínimos locais durante o treinamento [13]. Na prática, no entanto, ficou demonstrado que isso não seria um problema [9]. Outro problema era o fato de que a inferência de redes neurais demanda grande complexidade computacional, e as implementações existentes não eram muito eficientes. Operações de convolução são especialmente custosas, tornando as redes convolucionais ainda mais lentas que as redes neurais tradicionais [10]. A implementação eficiente de redes neurais convolucionais utilizando *graphical processing units* (GPUs, processadores especializados em operações matriciais capazes de realizar muitas operações paralelas) em 2006 por Chellapilla, Puri e Simard [14] foi de significativa importância no enorme desenvolvimento das CNNs que se seguiu.

No período entre 2006 e 2012, redes convolucionais começaram a se popularizar, e foram aplicadas a diversas tarefas, como classificação e reconhecimento de imagens e objetos (vide seção 4.3 de [10]), detecção facial [15] e segmentação de imagens [16]. Em 2012, Krizhevsky et al. [17] publicaram uma nova arquitetura, AlexNet, com um desempenho inédito na ImageNet que fez as redes convolucionais explodirem em popularidade e se tornarem a principal abordagem para visão computacional.

ImageNet é um dataset contendo mais de 14 milhões de imagens de alta resolução classificadas em mais de 20.000 classes [18]. As imagens foram coletadas na internet e classificadas por humanos. Desde 2010, tem sido realizada todos os anos a ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), uma competição para classificar imagens de um subconjunto da ImageNet, organizadas em aproximadamente 1000 classes com cerca de 1000 imagens cada. O dataset de treinamento contém aproximadamente 1,2 milhão de imagens, enquanto o conjunto de teste contém 150.000 imagens. As métricas usualmente utilizadas para avaliar o desempenho de um modelo na classificação de imagens da ImageNet são os erros top-5 e top-1, calculados como a fração das imagens do conjunto de teste nas quais o modelo não incluiu a classe correta entre as cinco, e uma, respectivamente, classificações dadas como mais prováveis. Na ILSVRC 2012, Krizhevsky et al., venceram a competição utilizando um *ensemble* de 7 versões de sua arquitetura, alcançando um erro top-5 de apenas 15,3%. Para comparação, nota-se que o segundo colocado conseguiu um erro top-5 de 26,2%, e o primeiro colocado na ILSVRC 2011, detentor do recorde até então, alcançou erro top-5 de 25,8% [18].

O sucesso da AlexNet se deveu à implementação de diversas novidades. As funções de ativação utilizadas, tradicionalmente funções sigmoide como a tangente hiperbólica, foram substituídas pela ReLU [19, 20], o que diminuiu o tempo de treinamento. A arquitetura desenvolvida também era mais larga e mais profunda, de modo que foi necessário dividir a rede entre duas GPUs. O aumento na quantidade de parâmetros da rede, no entanto, aumenta também sua capacidade para *overfitting*. Para contornar esse problema, duas técnicas foram

utilizadas. Primeiro, os autores aplicaram *Dropout*, uma técnica introduzida por Hinton et al. em [21] inicialmente para redes completamente conectadas, que consiste em, durante o treinamento, permitir que cada neurônio da rede seja omitido com uma certa probabilidade, neste caso 0,5; durante a fase de teste, todos os neurônios são utilizados, embora a saída de cada um seja multiplicada pelo valor da probabilidade, para compensar pelo maior número de neurônios ativos. Segundo, foi utilizado *data augmentation*, aplicando-se translações, reflexões horizontais, perturbações nas intensidades dos canais dos pixels e realizando análise de componente principal (PCA) nos valores dos pixels nas imagens do dataset.

Em 2014, Simonyan e Zisserman buscavam melhorar a acurácia das redes convolucionais aumentando sua profundidade. Eles desenvolveram uma arquitetura com 19 camadas, chamada de VGG [22], um significativo aumento na profundidade desde a AlexNet. Para limitar o aumento na complexidade, os autores mostraram que pilhas de convoluções 3×3 podem simular convoluções maiores, ao mesmo tempo utilizando menos parâmetros, adicionando regularização e incorporando mais camadas de funções de ativação, melhorando a função de decisão. A arquitetura resultante conseguiu o segundo lugar na ILSVRC 2014 com um ensemble de 7 modelos, com erro top-5 de 7,3%. Apesar das medidas para conter a complexidade da arquitetura, os modelos desenvolvidos por Simonyan e Zisserman contêm entre 133 milhões e 144 milhões de parâmetros.

Simultaneamente, Szegedy et al. [23] também buscavam aumentar a profundidade e a largura das redes convolucionais, no entanto evitando o aumento demasiado da complexidade computacional que tende a vir com esse objetivo, como observado com a VGG. Os autores introduziram o módulo Inception para CNNs, utilizado para a construção de redes Inception. Esse módulo utiliza filtros de tamanho variável para detectar padrões de diferentes tamanhos. Para reduzir a complexidade, o módulo utiliza convoluções 1×1 , como modo de redução de dimensionalidade antes de convoluções 3×3 e 5×5 , que são mais custosas. Uma das redes Inception, GoogLeNet, com 22 camadas, venceu a ILSVRC 2014, alcançando erro top-5 de 6,67%. Em contraste com a VGG, a GoogLeNet tem apenas cerca de 6,8 milhões de parâmetros.

Em 2015, He et al. [24] foram ainda mais longe, desenvolvendo uma arquitetura com 152 camadas de profundidade. No entanto, essa grande quantidade de camadas introduz um novo problema: os gradientes da função de perda em relação aos parâmetros das camadas mais rasas tende a 0, em um problema conhecido como *vanishing gradient*. Como o treinamento depende dos gradientes, torna-se difícil atingir convergência no modelo. Para remediar isso, a nova arquitetura, ResNet, introduziu o conceito de bloco residual. Em um bloco residual, ao final da transformação soma-se à saída as ativações que serviram de entrada para o bloco. Formalmente, denotando a transformação de x esperada do bloco por $\mathcal{H}(x)$, deixa-se

que o interior do bloco aprenda outra transformação $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$, e então soma-se \mathbf{x} à saída do bloco, obtendo ao final $\mathcal{F}(\mathbf{x}) + \mathbf{x} = \mathcal{H}(\mathbf{x})$. Efetivamente, há uma conexão entre camadas não consecutivas que “pula” o interior do bloco. Essa técnica permite a propagação do gradiente para camadas mais rasas, tornando o treinamento da rede mais fácil. Essa arquitetura venceu a ILSVRC 2015 com um ensemble, conseguindo erro top-5 de 3,57%.

2.5 Aplicações das redes convolucionais

Redes convolucionais têm aplicações em diversas áreas, como visão computacional, processamento de linguagem natural e reconhecimento de fala. A seguir, são apresentadas um pequeno apanhado de aplicações em visão computacional.

Em 2014, Li, Liu e Chan [25] mostraram que CNNs podem ser utilizadas para estimar a pose de corpos humanos, identificando partes do corpo e juntas entre elas. Também em 2014, Girshick et al. [26] conseguiram melhorias expressivas na tarefa de detecção de objetos utilizando redes convolucionais. Em 2015, Farfade, Saberian e Li [27] descreveram o uso de redes convolucionais para a detecção de rostos. Enquanto abordagens de ponta anteriores dependiam de inúmeros modelos treinados para lidar com diferentes poses, os autores mostraram que um único modelo convolucional era capaz de realizar a detecção.

Em 2016, Liang et al. [28] demonstraram que redes convolucionais são capazes de detectar câncer pulmonar em tomografias computadorizadas, incluindo instâncias que não foram detectadas por radiologistas. Kooi et al. [29] empregaram CNNs para ler mamografias, e obtiveram resultados comparáveis ou melhores que o estado da arte até então. Em 2018, foi demonstrado em um trabalho por Kermay et al. [6] a capacidade por redes convolucionais de detectar pneumonia em radiografias torácicas de crianças. Mais recentemente, em 2020, Chouhan et al. [30] alcançaram 96,4% de acurácia na detecção de pneumonia em radiografias utilizando um *ensemble* de CNNs.

2.6 Técnicas de explicabilidade

Com a tendência das redes convolucionais a ficarem mais e mais profundas e complexas, o problema da interpretabilidade dos resultados obtidos por elas ganhou importância. Redes neurais e convolucionais não têm uma estrutura naturalmente interpretável como regressões lineares e árvores de decisão, querendo o desenvolvimento de técnicas novas.

2.6.1 CAM

Class Activation Mapping (CAM) [31] é uma técnica que utiliza os mapas de *features* (ou ativações) de uma camada convolucional no final de uma CNN para gerar um mapa de calor para uma imagem sendo classificada, indicando regiões de importância para a classificação. CAM tem a limitação de ser compatível somente com redes completamente convolucionais. A última camada do classificador produz a probabilidade de cada classe c , aplicando a função softmax aos *scores* y^c de cada classe. Os scores são obtidos diretamente dos mapas de ativação da última camada convolucional, através soma ponderada do *global average pooling* de cada mapa. Se $A^k = (A_{i,j}^k)$ é o k -ésimo mapa de ativações da última camada convolucional, com pixels indexados por i, j , então o score y^c é obtido por

$$y^c = \sum_k w_k^c \sum_i \sum_j A_{i,j}^k,$$

onde w_k^c é o peso do k -ésimo mapa para a classe c , e é um parâmetro aprendido. O mapa de calor M^c produzido pelo CAM para a classe c é então obtido por

$$M^c = \sum_k w_k^c A_k.$$

2.6.2 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [32] é uma generalização do CAM que pode ser aplicada à uma classe maior de redes convolucionais. Assim como CAM, o Grad-CAM cria um mapa de calor para a imagem sendo classificada, destacando partes de importância. Diferentemente do CAM, o Grad-CAM não requer que a rede seja inteiramente convolucional, permitindo que ela contenha camadas completamente conectadas. Isso dispensa a necessidade de retreinar um modelo sem suas camadas completamente conectadas. A técnica cria uma visualização levando em consideração os gradientes em relação a uma classe da saída da rede

Denotando por y^c o score para uma classe c antes de aplicar a função softmax, o Grad-CAM utiliza os gradientes $\frac{\partial y^c}{\partial A^k}$ de y^c com relação às ativações dos mapas de *features* A^k da última camada convolucional. O mapa de calor L^c com relação à classe c é dado por uma combinação linear dos mapas A^k :

$$L^c := ReLU \left(\sum_{k=1}^K \alpha_k^c A^k \right),$$

onde o peso α_k^c do mapa k é obtido por *global average pooling* dos gradientes desse mapa:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad (1)$$

onde Z é o número de pixels em cada mapa, e i, j iteram as duas dimensões de cada mapa. A utilização da função ReLU na combinação linear restringe a visualização às *features* que contribuem de maneira positiva para a classe sendo examinada.

A Figura 4 ilustra essa técnica, juntamente com suas aplicações em outras áreas.

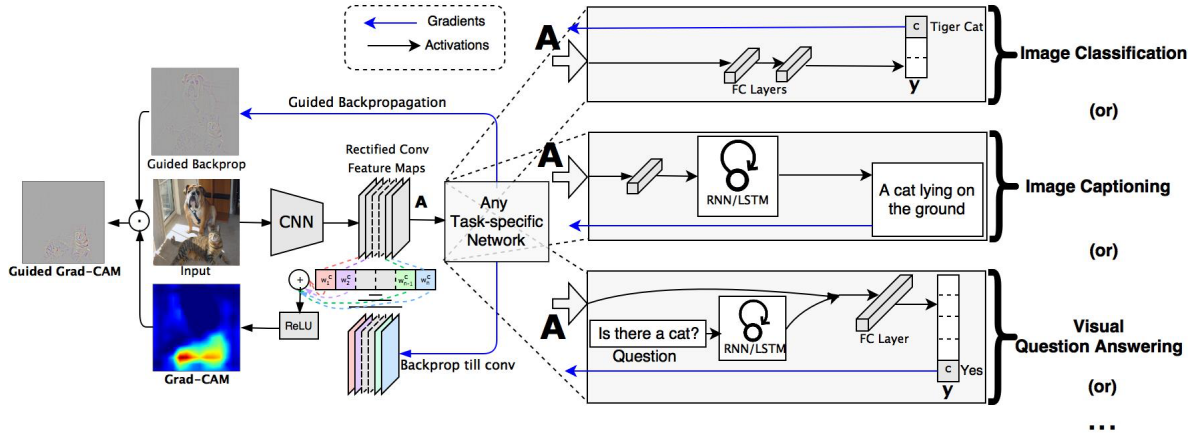


Figura 4: Esquemática do Grad-CAM. Figura extraída de [32]. © 2017 IEEE.

2.6.3 Grad-CAM++

Grad-CAM++ [33] se baseia no Grad-CAM. Essa técnica busca solucionar alguns problemas com a antecessora. Especificamente, duas deficiências do Grad-CAM são a dificuldade para lidar com a localização de objetos que têm múltiplas ocorrências na imagem e a localização de apenas partes de objetos em vez do todo, devido ao *global average pooling* das derivadas parciais, sem ponderação. Grad-CAM++ utiliza uma média ponderada das derivadas parciais do score y^c após a função softmax em relação aos mapas de *features* para calcular sua importância na combinação linear de todos os mapas de *features*. A equação (1) se torna então

$$\alpha_k^c = \sum_i \sum_j \beta_{ij}^{kc} \frac{\partial y^c}{\partial A_{ij}^k},$$

onde os coeficientes β_{ij}^{kc} para cada pixel i, j da camada k são dados por

$$\beta_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}}{2 \frac{\partial^2 y^c}{(\partial A_{ij}^k)^2} + \sum_{i'} \sum_{j'} A_{i'j'}^k \left[\frac{\partial^3 y^c}{(\partial A_{ij}^k)^3} \right]}.$$

2.6.4 LIME

Local Interpretable Model-agnostic Explanations (LIME) [34] é uma técnica da explicabilidade desenvolvida para ser utilizada com qualquer classe de modelos de classificação. A ideia por trás dessa técnica consiste em ajustar outro modelo, simples e interpretável, que forneça uma boa aproximação do modelo classificador que se deseja explicar na vizinhança de uma amostra.

Dada uma amostra x a ser classificada (por exemplo, um texto, imagem ou dados categóricos) e um vetor binário $x' \in \{0, 1\}^{d'}$ que representa x através a presença ou ausência de componentes (e.g. palavras, super-pixels etc.), LIME constrói uma vizinhança de x' , com novos vetores obtidos pela substituição de quantidades aleatórias de componentes não nulos de x' (i.e., com valor 1) por 0. Esses vetores são mapeados a amostras na vizinhança de x em seu espaço original, que correspondem a transformações de x obtidas removendo componentes da amostra. A predição para essas amostras, obtidas utilizando o modelo a ser examinado, são então transferidas aos respectivos vetores da vizinhança de x' , e ajusta-se um modelo interpretável (como um modelo linear ou árvore de decisão) a esse pequeno dataset. A interpretação desse modelo é utilizada para explicar a influência dos componentes na decisão do modelo.

Para classificadores de imagens, os componentes utilizados são regiões contíguas de pixels similares por alguma métrica, chamadas de super-pixels.

2.7 Considerações finais

Nesta seção, foram apresentados um resumo da fundamentação teórica necessária para a compreensão do trabalho: as redes neurais artificiais e redes neurais convolucionais, utilizadas com desempenho de ponta na tarefa de classificação de imagens, o desenvolvimento das arquiteturas de redes convolucionais profundas guiado pela busca por maior desempenho e menor complexidade, e técnicas desenvolvidas para a interpretação dos resultados obtidos por CNNs. Na próxima seção, serão introduzidos trabalhos relacionados que empregaram redes convolucionais para detectar Covid-19 em imagens médicas.

3 Trabalhos relacionados

Ao longo do ano, com o avanço da pandemia de Covid-19, foi proposto o uso de imagiologia média (radiografias e tomografias computadorizadas) para auxiliar na detecção de Covid-19 em pacientes. Isso foi seguido por trabalhos propondo a utilização de técnicas de visão computacional para a automatização dessa tarefa. Os trabalhos podem ser divididos entre aqueles que utilizam radiografias torácicas e alguns que utilizam outros métodos de imagiologia, nomeadamente tomografias computadorizadas e ultrassom.

3.1 Radiografias torácicas

Alguns trabalhos se colocaram a comparar o desempenho de arquiteturas com *fine-tuning* de redes convolucionais já estabelecidas na tarefa de detecção de Covid-19. Apostolopoulos e Bessiana [35] aplicaram *transfer learning* para treinar modelos VGG19, MobileNet, Inception, Xception e Inception-ResNet-v2 tanto em classificação binária e multiclasse (Covid-19, Normal, Pneumonia), utilizando radiografias coletadas do *COVID-19 image data collection* e outros dois datasets. Reportaram 97,82% de acurácia. Kassani et al. [36] utilizaram as arquiteturas de diversas CNNs (DenseNet, ResNet, Inception-v3, Inception-Resnet-v2, VGGNet, NASNet e MobileNet) para realizar *feature extraction* de radiografias, e então empregaram diversos classificadores clássicos de *machine learning* para realizar classificação binária das imagens. Reportaram 99% de acurácia utilizando DenseNet121 com classificador Bagging Tree, utilizando o dataset *COVID-19 image data collection*.

Outros trabalhos utilizaram as arquiteturas já conhecidas como ponto de partida para a utilização de outras técnicas. Para lidar com o problema da pequena quantidade de radiografias de pacientes com Covid-19 disponíveis, Oh, Park e Ye [37] propuseram uma abordagem baseada na utilização de diversos fragmentos (ou *patches*) das radiografias com o pulmão segmentado. Adicionalmente, o modelo também foi treinado em radiografias completas. O sistema de fragmentos funciona selecionando, de maneira aleatória, 100 fragmentos quadrados de 224×224 pixels da imagem sobre o pulmão, aplicando o classificador e então tomando o voto majoritário dos fragmentos. O modelo classifica as radiografias entre as classes Normal, Covid-19 + Pneumonia Viral, Tuberculose e Pneumonia Bacterial. O classificador utilizado foi a ResNet-18, e o modelo alcançou 70,7% de acurácia utilizando a imagem completa e 88,9% de acurácia com a abordagem utilizando fragmentos. Também propuseram um novo método de interpretabilidade baseado no Grad-CAM, utilizando o resultado do Grad-CAM em cada um dos fragmentos. Punj e Agarwal [38] substituíram a função de perda padrão de arquiteturas por uma com classes ponderadas (*weighted class loss function*), e utilizaram *oversampling* aleatório e *transfer learning* com arquiteturas como ResNet, Inception-v3,

Inception-ResNet-v2, DenseNet169 e NASNetLarge para classificar radiografias torácicas. Avaliaram tanto classificação binária quanto classificação multiclasse, entre Normal, Covid-19 e Pneumonia, empregando LIME para examinar a interpretabilidade. O dataset foi obtido combinando o *COVID-19 image data collection* (uma versão agora desatualizada) e o dataset disponibilizado pela Sociedade Radiológica da América do Norte, que faz parte do dataset ChestX-ray14 de radiografias. O artigo reporta 98% de acurácia com a NASNetLarge, utilizando oversampling. Zhang et al. [39] propuseram utilizar um modelo profundo de detecção de anomalias para classificação binária de radiografias torácicas. O modelo possui uma rede convolucional para a extração de *features* de alto nível das radiografias e dois componentes para realizar a classificação e a detecção de anomalias. A CNN utilizada é a ResNet18, e os componentes de classificação e detecção de anomalias são ambos compostos por uma camada convolucional seguido de um *multi-layer perceptron*. A classificação depende de um *threshold*: com 0,5, obtiveram sensibilidade de 72% e especificidade de 97,97%; com 0,15, eles foram 96% e 70,65%, respectivamente. Os autores utilizaram Grad-CAM para visualizar as ativações. Goshal e Tucker [40] investigaram o uso de redes neurais convolucionais Bayesianas (BCNN) baseadas em *Dropweights* para estimar incerteza na classificação de radiografias. Diferencia-se por procurar uma maneira de permitir que o modelo diga que “não sabe a resposta”. Os autores utilizaram a arquitetura ResNet50V2 para a classificação, e escolheram uma função de custo assimétrica, para dar maior ênfase a resultados falso negativos. Também utilizaram vários métodos para avaliar a interpretabilidade: mapas de saliência, CAM, Grad-CAM e Guided Backpropagation. Li et al. [41] propuseram a utilização de *discriminative cost-sensitive learning* para a classificação multiclasse (Covid-19, Normal, Outro). O trabalho aborda o fato de que a diferença entre radiografias de diferentes tipos de pneumonia não é muito grande, aumentando o custo de erroneamente classificar casos de Covid-19 em outras classes. Utilizaram a VGG16, introduzindo algumas modificações, para a classificação.

Luz et al. [42] notam que as arquiteturas utilizadas são em geral bastante complexas, impróprias para implantação em locais sem grandes recursos computacionais. Os autores utilizaram uma família de modelos mais leve, EfficientNet, para realizar classificação multiclasse (Normal, Covid-19, Outra Pneumonia). Utilizaram classificação hierárquica, e aplicaram CAM para visualizar as ativações.

Alguns grupos propuseram novas arquiteturas. Wang, Lin e Wong [43] introduziram uma nova rede convolucional para a realização de classificação multiclasse (Normal, Covid-19, Outra Infecção) de radiografias torácicas, nomeada COVID-Net, utilizando conexões de longo alcance e blocos projeção-expansão-projeção (blocos PEPX). Reportaram ter alcançado acurácia de 93,3% e sensibilidade de 91% para a classe Covid-19. Aplicaram GSInquire para examinar a explicabilidade do modelo. Baseando-se na DarkNet, modelo

já estabelecido, Ozturk et al. [44] propõe uma arquitetura com 17 camadas convolucionais, chamada DarkCovidNet, para a detecção de Covid-19 em radiografias torácicas. O modelo foi empregado tanto em classificação binária quanto em um cenário multiclasse. A rede convolucional foi treinada desde o princípio, sem o emprego de *transfer learning*, e os autores reportaram acurácia de 98,08% no caso binário e 87,02% no caso multiclasse. O dataset utilizado para o treinamento é uma combinação de imagens de raios-X de Covid-19 provenientes (de uma versão agora antiga) do *COVID-19 image data collection* e de raios-X normais e com pneumonia do dataset ChestX-ray8. O dataset foi disponibilizado para reprodução.

3.2 Tomografias computadorizadas e ultrassom

Maghdid et al. [45] aplicaram *transfer learning* para AlexNet e propuseram uma pequena rede convolucional com somente uma camada convolucional para a detecção de Covid-19 em radiografias torácicas e tomografias computadorizadas com classificação binária. Observaram acurácia de 94% e 98% aplicando a arquitetura própria e a AlexNet, respectivamente, em radiografias, e 94,1% e 82%, respectivamente, para tomografias. Born et al. [46] criaram um dataset utilizando vídeos de ultrassom de pacientes com Covid-19, pneumonia bacterial e saudáveis, e propuseram uma arquitetura para rede convolucional baseada na VGG-16 para realizar a classificação dos *frames* e dos vídeos como um todo. Os autores reportaram acurácia de 89% na classificação das imagens individuais e 92% na classificação dos vídeos por voto majoritário.

3.3 Considerações finais

Esta seção apresentou diversos trabalhos encontrados na literatura que abordam a detecção de Covid-19 em vários tipos de imagens médicas. Primeiro, foram descritas diversas técnicas aplicadas à classificação de radiografias torácicas, incluindo o emprego de classificadores baseados em redes convolucionais, a classificação por modelos não-neurais de *features* extraídas por CNNs e a proposta de algumas arquiteturas novas para realizar a classificação. Também foram mencionados trabalhos que utilizam tomografias computadorizadas e ultrassom para realizar a detecção de Covid-19.

4 Metodologia

O desenvolvimento do trabalho passa pelas etapas a seguir. Em primeiro lugar, um dataset com radiografias de todas as classes é montado, com considerações como origem das imagens e balanço na quantidade de imagens de cada classe. Outro dataset é derivado do primeiro, excluindo das imagens a região fora dos pulmões. Então, seleciona-se as arquiteturas a serem utilizadas, e dá-se o processo de treinamento dos modelos. Finalmente, as métricas dos modelos obtidos são analisadas, e as explicações são extraídas e analisadas.

4.1 Construção do dataset

A construção e utilização do dataset consistiu na coleta das imagens, aplicação de *data augmentation* durante o treinamento e realização de *segmentação* das radiografias.

4.1.1 Datasets disponíveis

Há alguns datasets de radiografias torácicas disponíveis. A seguir, são listados datasets que não disponibilizam dados para Covid-19.

O **NIH Chest X-ray Dataset** [47] (ou **ChestX-ray14**, previamente **ChestX-ray8**) é uma coleção de 112.120 radiografias disponibilizadas pelo *National Institutes of Health*, distribuídas em 14 categorias para doenças e uma sem achados clínicos.

Mendeley disponibiliza o **Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images** [48], com 5.856 radiografias, classificadas como Normal, Pneumonia Bacteriana ou Pneumonia Viral. As radiografias foram coletadas de crianças, de modo que a utilização desse dataset em conjunto com outros pode ser problemática, devido a possíveis diferenças morfológicas entre os pacientes nos datasets.

O **JSRT Database** [49] é fornecido pela *Japanese Society of Radiological Technology*, e contém 247 radiografias. Em conjunto com o **SCR database** [50], que contém segmentações para todas as 247 imagens, pode ser utilizado para treinar modelos para segmentação dos pulmões. O trabalho de Oh et al. [37], reproduzido neste trabalho, utilizou estes datasets para treinar um segmentador.

A *U.S. National Library of Medicine* fornece dois datasets com radiografias normais normais e com tuberculose [51]: **Montgomery County X-ray Set**, com 138 radiografias, e **Shenzhen Hospital X-ray Set**, com 662 radiografias.

O *Stanford Machine Learning Group* disponibiliza o dataset **CheXpert** [52]. Esse dataset consiste de 224.316 radiografias, obtidas de 65.240 pacientes, acompanhadas de rótulos indicando a presença de 14 observações comuns em radiografias do tórax.

A seguir, são descritos datasets que contém radiografias de pacientes com Covid-19. Devido à pequena quantidade de imagens disponíveis, há grande sobreposição entre os datasets.

O **COVIDx Dataset** [43] compila radiografias com Covid-19 de diversas fontes. O dataset contém 14.258 radiografias distribuídas entre as categorias Normal, Pneumonia e Covid-19. Entre as radiografias, 619 são da classe Covid-19.

O **COVID-19 image data collection** [53] também compila radiografias obtidas de diversas fontes para a detecção de Covid-19. O dataset conta com 654 radiografias distribuídas em diversas categorias hierarquicamente, das quais 451 são da classe Covid-19.

O **CoronaHack Dataset** [54] reúne 5933 radiografias, com 58 de Covid-19 colhidas do *COVID-19 image data collection*. A maioria das radiografias é oriunda do dataset disponibilizado por Mendeley. O **COVID-19 X rays Dataset** [55] contém 78 radiografias, das quais 71 são de Covid-19, bem como 16 tomografias de Covid-19.

4.1.2 Data Augmentation

Redes neurais, em particular redes convolucionais, requerem grandes quantidades de dados durante o treinamento para alcançar um bom desempenho e conseguir generalizar o resultado para dados fora do conjunto de treinamento.

Dada a novidade da Covid-19, não houve tempo ainda para que fossem criados datasets massivos de radiografias de pacientes acometidos da doença, como há para outros tipos de doenças pulmonares e para pulmões saudáveis, criando dois empecilhos para o problema de classificação. Em primeiro lugar, a quantidade de radiografias disponíveis pode não ser suficiente para ensinar uma rede convolucional a detectar Covid-19 de maneira confiável. Segundo, para as outras classes, é necessário escolher entre utilizar uma grande quantidade entre as radiografias disponíveis ou escolher uma quantidade semelhante àquela para Covid-19. A primeira opção cria um dataset com classes altamente desbalanceadas que pode levar a *overfitting*, enquanto a segunda limita a quantidade de dados no conjunto de treinamento e pode limitar o desempenho do classificador.

Uma maneira de combater esses problemas é a utilização de *data augmentation*. Essa é uma técnica para ampliar o dataset utilizado criando novos dados através da aplicação de transformações bem comportadas, possivelmente aleatórias, aos dados existentes. As novas amostras podem ser armazenadas em disco, efetivamente aumentando o número de amostras no dataset, ou as transformações podem ser aplicadas com alguma probabilidade aos dados existentes durante o treinamento. As transformações comumente utilizadas para aumentar datasets de imagens incluem reflexão horizontal (e vertical, com menor frequência), rotação,

recorte da imagem, cisalhamento, alteração do contraste, perturbação das cores dos pixels e ocultação de partes da imagem.

Em nosso dataset utilizamos *data augmentation*, aplicando as transformações durante o treinamento.

4.1.3 Dataset com segmentação dos pulmões

A segmentação de uma imagem é a atribuição de um rótulo para cada pixel da imagem, dividindo-a em segmentos, ou conjuntos de pixels que compartilham algumas características. Uma segmentação pode, por exemplo, especificar a localização de objetos distintos na imagem, separar o fundo de uma fotografia da parte em primeiro plano, ou dividir uma fotografia de satélite de uma área rural por tipo de plantação. Essa segmentação pode ser feita de maneira manual ou inferida por um modelo especializado.

Neste trabalho, é de especial interesse as partes das radiografias onde estão os pulmões. Como pneumonias são doenças pulmonares, é natural a expectativa de que modelos classificadores precisem somente dos pulmões para realizar uma predição, ignorando o restante da radiografia. Assim como [37], também treinamos classificadores que recebem imagens com os pulmões segmentados, isto é, que contém os pixels originais onde os pulmões estão localizados e tiveram o restante substituído por pixels pretos. A Figura 5 apresenta um exemplo de radiografia segmentada.



Figura 5: Exemplo de radiografia segmentada

Para esses classificadores, uma versão do dataset utilizado, com os pulmões segmentados, foi criada utilizando o modelo de segmentação por Selvan et al. em [56].

4.2 Treinamento dos modelos classificadores

Para obter as classificações e examinar o funcionamento dos modelos, algumas arquiteturas de redes convolucionais foram selecionadas e treinadas.

Foram escolhidas arquiteturas bem conhecidas, ResNet-18, ResNet-152, VGG-16, GoogLeNet, DenseNet-121, e AlexNet, por sua comprovada eficácia aplicadas à ImageNet e disponibilidade de implementações e parâmetros pré-treinados. Dentre os trabalhos relacionados, dois modelos com implementação disponível foram escolhidos para reprodução e inspeção. Um, de Oh, Park e Ye [37], que propôs a classificação com ResNet-18 de diversos fragmentos das imagens com os pulmões segmentados, limitando a visão do modelo. Outro, DarkCovidNet, que apresentou uma nova arquitetura baseada em outra, DarkNet, sendo treinada sem o uso de *fine tuning*.

A tarefa de projetar a arquitetura de uma rede neural não é trivial. Assim, empregar arquiteturas já bem estabelecidas é uma boa estratégia, praticada amplamente na criação de novos classificadores. Além da garantia de um bom *design*, os parâmetros de instâncias pré-treinadas dessas arquiteturas estão disponíveis, geralmente treinadas na ImageNet. Apesar de elas não serem imediatamente capazes de realizar previsões para novas classes em um novo domínio de imagens, o treinamento em um dataset com a complexidade da ImageNet significa que os modelos já são capazes de reconhecer *features* complexas em imagens. A utilização dos parâmetros obtidos com a ImageNet como ponto de partida, em vez de começar com parâmetros aleatórios, efetivamente aumenta a quantidade de dados utilizados para o treinamento, além de diminuir o tempo de treinamento do novo classificador. Essa técnica, chamada de *fine tuning*, é por conseguinte útil para o treinamento de modelos quando se tem uma pequena quantidade de dados disponível.

O treinamento foi implementado em Python, utilizando a biblioteca PyTorch.

4.3 Métodos para interpretabilidade

No contexto médico, onde o custo de erros são altos, a explicação de um resultado é tão importante quanto a classificação. Para avaliar o funcionamento dos modelos treinados e comparar as técnicas, foram escolhidas Grad-CAM e Grad-CAM++.

Como notado, a aplicação de Grad-CAM e Grad-CAM++ estima a importância dos pixels para a classificação através das ativações da última camada convolucional, ponderadas pela influência de cada mapa de ativações individual na classificação.

Devido à incompatibilidade do CAM com as arquiteturas utilizadas, essa técnica não foi utilizada.

4.4 Considerações finais

A seção apresentou datasets relevantes que estão disponíveis, o problema de desbalançamento de classes, presente devido à novidade da Covid-19, as técnicas de *data augmentation* e segmentação e as arquiteturas de classificadores a serem treinados.

5 Resultados

Os classificadores treinados foram avaliados utilizando as seguintes métricas: acurácia, precisão, revocação (*recall*) e valor F (*F1 Score*). Para entender os erros realizados pelos modelos, as matrizes de confusão para as classificações foram examinadas. Finalmente, a fim de investigar o funcionamento interno dos modelos, foram aplicadas as técnicas de Grad-CAM e Grad-CAM++.

5.1 Avaliação de métricas de classificação

A acurácia de um classificador é definida como a razão entre a quantidade de classificações realizadas corretamente e a quantidade total de classificações, ou seja, é a probabilidade observada de que o classificador tome a decisão correta.

A precisão é calculada para cada classe separadamente. Essa métrica é definida pela razão entre a quantidade de classificações corretas em uma classe e a quantidade de predições para aquela classe, isto é, pela proporção de detecções positivas que realmente correspondem à classe.

Assim como a precisão, a revocação é calculada para cada classe individualmente, e é definida pela razão entre a quantidade de classificações realizadas corretamente para uma classe e a quantidade total de elementos naquela classe. Dessa maneira, corresponde à proporção dos elementos daquela classe que foram corretamente identificados.

O valor F combina ambos precisão e revocação, sendo definido pela média harmônica entre eles,

$$F = 2 \cdot \frac{\text{precisão} \cdot \text{revocação}}{\text{precisão} + \text{revocação}}.$$

O valor F varia entre 0 (pior) e 1 (melhor).

5.1.1 Acurácia

A Tabela 1 exibe a acurácia alcançada pelos modelos treinados utilizando cada arquitetura de redes convolucionais, separados pela maneira como as radiografias foram utilizadas durante o treinamento. A acurácia se refere ao conjunto de testes do dataset, não utilizado durante o treinamento.

Observa-se que a DenseNet-121 obteve a melhor acurácia entre as arquiteturas, com 92,8%, seguida igualmente pela AlexNet e ResNet-121. Note-se que a AlexNet é uma das arquiteturas de redes convolucionais mais simples. A situação é a mesma observando somente os modelos treinados com radiografias integrais. Com treinamento utilizando as radiografias

| Arquitetura | Acurácia | |
|--------------|-------------------|---------------------|
| | Imagens integrais | Imagens segmentadas |
| AlexNet | 0,910 | 0,802 |
| DenseNet-121 | 0,928 | 0,844 |
| GoogLeNet | 0,862 | 0,808 |
| ResNet-18 | 0,886 | 0,862 |
| ResNet-152 | 0,910 | 0,820 |
| VGG-16 | 0,886 | 0,796 |
| DarkCovidNet | 0,820 | 0,623 |
| Oh et al. | — | 0,904 |

Tabela 1: Acurácia para cada modelo treinado

segmentadas, o modelo de Oh et al. conseguiu a melhor acurácia, com 90,4%, seguido da ResNet-18, com 86,2%.

É importante ressaltar aqui que a acurácia dos modelos treinados com as radiografias segmentadas foi sempre menor que a dos modelos com as respectivas arquiteturas treinados utilizando as radiografias integrais, chegando a uma redução absoluta de mais de 10% no caso da AlexNet. Esse deterioramento fornece evidência de que havia informações indesejadas relevantes para a detecção pelos modelos presentes nas regiões das radiografias externas ao pulmão, removidas após a segmentação.

5.1.2 Precisão

Na Tabela 2, estão listadas as precisões observadas de cada arquitetura para cada classe, separadas com relação às radiografias integrais e segmentadas. As precisões reportadas foram calculadas utilizando o conjunto de testes.

A análise nos modelos que utilizam as radiografias integrais mostra uma grande precisão na detecção de radiografias normais em quase todos os modelos, com um valor levemente menor para a AlexNet. As detecções de Covid-19 costumam estar corretas para três dos modelos, com precisão igual a ou maior que 97,5%. Outras arquiteturas têm um desempenho bom, mas não tão alto. A detecção de pneumonia, no entanto, não alcança níveis tão altos, embora sejam razoáveis.

Para as radiografias segmentadas, a precisão na detecção de radiografias normais é grande, acima de 90% para a maioria dos modelos testados. As precisões para as classes de Covid-19 e pneumonia, no entanto, são mais pobres, com a exceção do modelo de Oh et al., que conseguiu 95,3% de precisão para Covid-19 e uma marca menor, mas maior que as

| Arquitetura | Precisão | | | | | |
|--------------|-------------------|--------|-----------|---------------------|--------|-----------|
| | Imagens integrais | | | Imagens segmentadas | | |
| | Covid-19 | Normal | Pneumonia | Covid-19 | Normal | Pneumonia |
| AlexNet | 0,976 | 0,894 | 0,881 | 0,833 | 0,881 | 0,712 |
| DenseNet-121 | 0,975 | 0,938 | 0,887 | 0,784 | 0,921 | 0,811 |
| GoogLeNet | 0,824 | 0,930 | 0,831 | 0,833 | 0,885 | 0,718 |
| ResNet-18 | 0,857 | 0,933 | 0,862 | 0,860 | 0,950 | 0,781 |
| ResNet-152 | 0,976 | 0,935 | 0,844 | 0,750 | 0,962 | 0,758 |
| VGG-16 | 0,889 | 0,949 | 0,825 | 0,833 | 0,960 | 0,667 |
| DarkCovidNet | 0,830 | 0,857 | 0,772 | 0,610 | 0,692 | 0,557 |
| Oh et al. | — | | | 0,953 | 0,933 | 0,844 |

Tabela 2: Precisão para cada modelo treinado

outras, de 84,4% para pneumonia.

Assim como na acurácia, houve queda na precisão da classificação de radiografias segmentadas em relação às integrais para diversos modelos. AlexNet e DenseNet-121 perderam precisão para todas as classes. A precisão para certas classes de alguns modelos, no entanto, melhorou. GoogLeNet e ResNet-18 tiveram ganhos moderados na precisão da detecção de Covid-19, enquanto ResNet-18, ResNet-152 e VGG-16 melhoraram a precisão para radiografias normais. A precisão para pneumonia diminuiu expressivamente para todos os modelos.

Embora alguns modelos tenham melhorado a precisão para uma dupla de classes, a maioria das precisões sofreu com a remoção do conteúdo externo ao pulmão das radiografias, indicando novamente sua relevância na classificação.

5.1.3 Revocação (*recall*)

A Tabela 3 apresenta a revocação obtida para as classes de cada modelo treinado, comparando os treinamentos com imagens integrais e segmentadas.

Os dados mostram que os classificadores treinados com as radiografias integrais são capazes de encontrar a vasta maioria das radiografias de pacientes com Covid-19, exibindo revocações superiores a 90%. AlexNet e DenseNet-121 também capturaram mais de 90% das radiografias normais, mas o restante das arquiteturas tiveram menores revocações, ainda que razoáveis, nessa classe. As revocações para a classe de pneumonia apresentaram resultados similares, com dois modelos, DenseNet-121 e ResNet-152, acima de 90%.

No caso dos classificadores para imagens segmentadas, somente três modelos foram capazes de identificar mais que 90% das radiografias da classe de Covid-19. Os modelos

| Arquitetura | Revocação | | | | | |
|--------------|-------------------|--------|-----------|---------------------|--------|-----------|
| | Imagens integrais | | | Imagens segmentadas | | |
| | Covid-19 | Normal | Pneumonia | Covid-19 | Normal | Pneumonia |
| AlexNet | 0,953 | 0,908 | 0,881 | 0,814 | 0,800 | 0,797 |
| DenseNet-121 | 0,907 | 0,938 | 0,932 | 0,930 | 0,892 | 0,729 |
| GoogLeNet | 0,977 | 0,815 | 0,831 | 0,814 | 0,831 | 0,780 |
| ResNet-18 | 0,977 | 0,862 | 0,847 | 0,860 | 0,877 | 0,847 |
| ResNet-152 | 0,930 | 0,892 | 0,915 | 0,907 | 0,785 | 0,797 |
| VGG-16 | 0,930 | 0,862 | 0,881 | 0,814 | 0,738 | 0,847 |
| DarkCovidNet | 0,907 | 0,831 | 0,746 | 0,581 | 0,692 | 0,576 |
| Oh et al. | — | | | 0,953 | 0,862 | 0,915 |

Tabela 3: Revocação para cada modelo treinado

restantes exibiram revocação razoável, acima de 80%. A revocação das radiografias normais ficou aproximadamente entre 80% e 90%, à exceção da VGG-16, com 73,8%. O modelo com revocação mais alta na classe de pneumonia é o de Oh et al., com 91,5%, enquanto os outros modelos ficaram limitados entre 73% e 85%, aproximadamente.

Espelhando os resultados da acurácia e precisão, todos, com exceção de um, os valores para revocação foram reduzidos nos classificadores treinados com radiografias segmentadas. A exceção ocorreu na ResNet-18, cuja revocação para as radiografias normais subiu levemente de 89,6% para 91,2%. Essa tendência dá lastro à hipótese de que os modelos se atentam a características alheias ao pulmão.

5.1.4 Valor F

A Tabela 4 informa o valor F dos classificadores para cada classe, novamente separando-os pelo tipo da radiografia. A coluna μ indica a média das classes.

Os modelos de imagem integral exibem boas pontuações para as classes Covid-19 e Normal, e valores um pouco menores para a classe de pneumonia. Mantendo a tendência, as pontuações para os classificadores de radiografias segmentadas são menores do que as correspondentes de radiografias integrais, e são menores para Pneumonia do que para as outras classes.

5.1.5 Matriz de Confusão

Nas Figuras 6 a 7 abaixo são apresentadas as matrizes de confusão para os modelos treinados, respectivamente, com radiografias integrais e segmentadas. As classes Covid-

| Arquitetura | Valor F | | | | | | | |
|--------------|-------------------|--------|--------|-------|---------------------|--------|--------|-------|
| | Imagens integrais | | | | Imagens segmentadas | | | |
| | Covid-19 | Normal | Pneum. | μ | Covid-19 | Normal | Pneum. | μ |
| AlexNet | 0,965 | 0,901 | 0,881 | 0,916 | 0,824 | 0,839 | 0,752 | 0,805 |
| DenseNet-121 | 0,940 | 0,938 | 0,909 | 0,929 | 0,851 | 0,906 | 0,768 | 0,842 |
| GoogLeNet | 0,894 | 0,869 | 0,831 | 0,864 | 0,824 | 0,857 | 0,748 | 0,810 |
| ResNet-18 | 0,913 | 0,896 | 0,855 | 0,888 | 0,860 | 0,912 | 0,813 | 0,862 |
| ResNet-152 | 0,952 | 0,913 | 0,878 | 0,915 | 0,821 | 0,864 | 0,777 | 0,821 |
| VGG-16 | 0,909 | 0,903 | 0,852 | 0,888 | 0,824 | 0,835 | 0,746 | 0,802 |
| DarkCovidNet | 0,867 | 0,844 | 0,759 | 0,823 | 0,595 | 0,692 | 0,567 | 0,618 |
| Oh et al. | — | | | | 0,953 | 0,896 | 0,878 | 0,909 |

Tabela 4: Valor F para cada modelo treinado

19, Normal e Pneumonia correspondem, respectivamente, às classes C, N e P nas figuras. As classes verdadeiras estão representadas à esquerda, enquanto as classes preditas ficam embaixo.

A análise das matrizes de confusão dos classificadores treinados com as radiografias integrais indica que a maioria dos erros desses modelos se concentram em dois deles: a classificação de radiografias da classe Pneumonia como radiografias normais; e a classificação de radiografias normais como da classe Pneumonia. Outros erros comuns acontecem com a classificação errada entre as classes Pneumonia e Covid-19, e erros de classificação entre as classes Covid-19 e Normal são raros.

Os modelos treinados a partir de radiografias segmentadas exibem um comportamento diferente. Os erros mais comuns se referem à classificação errada de radiografias normais como sendo da classe Pneumonia. A classificação errada recíproca, de imagens da classe Pneumonia como radiografias normais, não é tão comum. A classificação errada entre as classes Covid-19 e Pneumonia é responsável por quase todos os outros erros, e, como para as imagens integrais outra classe de modelos, erros entre Covid-19 e Normal são raros.

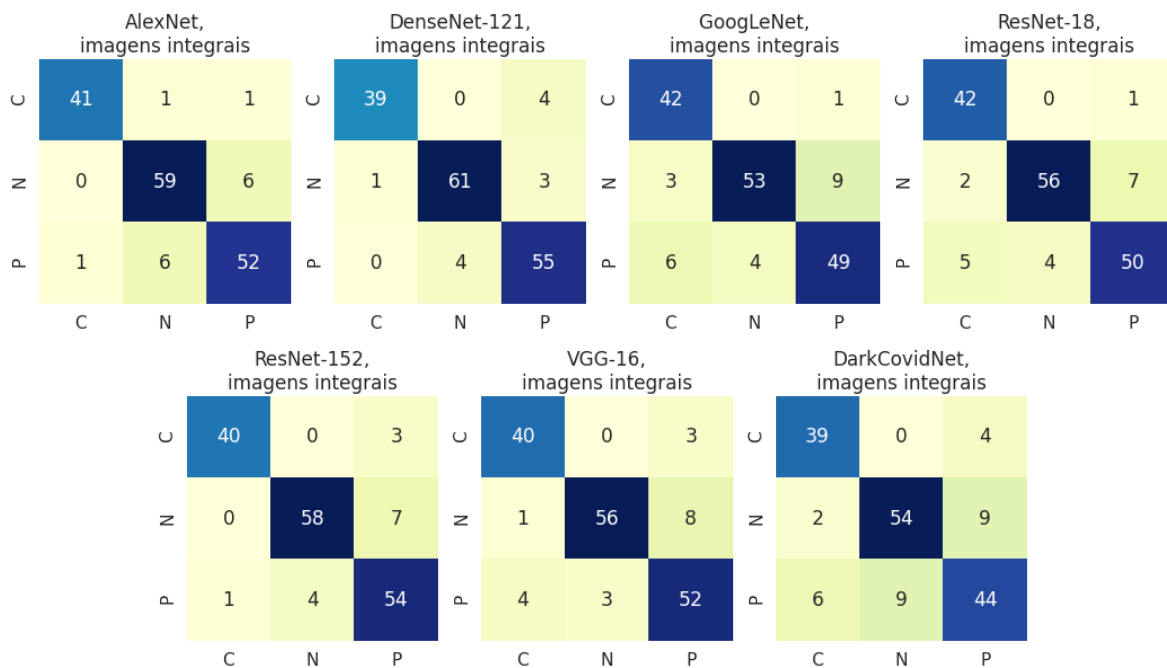


Figura 6: Matrizes de confusão para modelos treinados com radiografias integrais

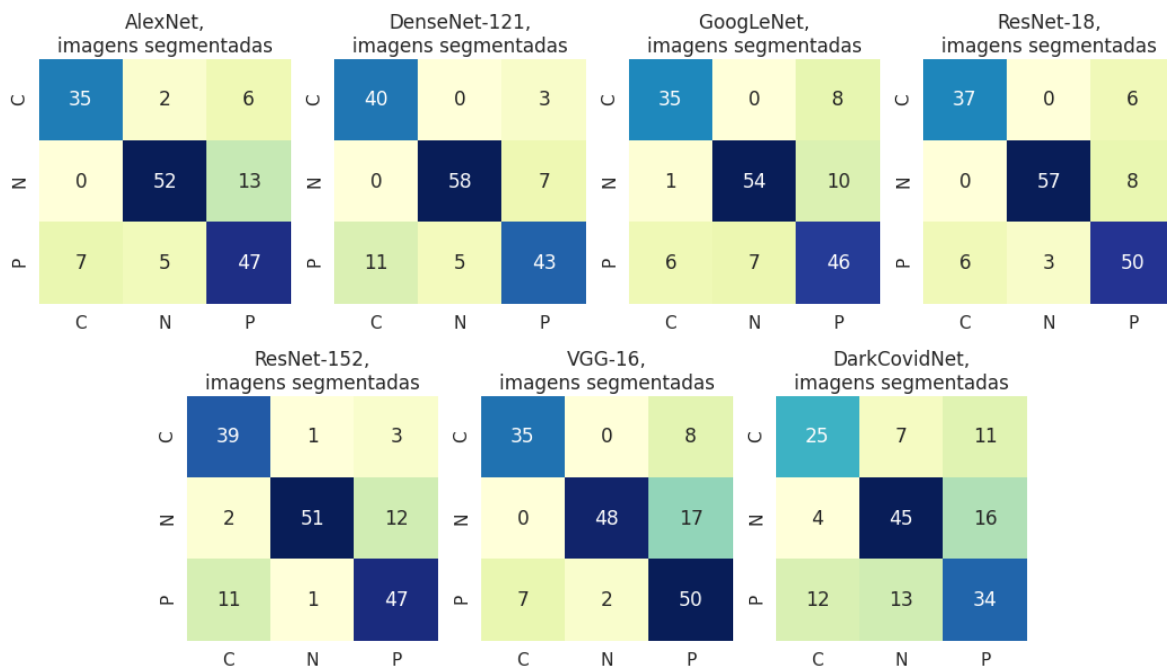


Figura 7: Matrizes de confusão para modelos treinados com radiografias integrais

5.2 Análise de interpretabilidade

As técnicas de Grad-CAM e Grad-CAM++ foram aplicadas nos modelos treinados para gerar explicações visuais das decisões tomadas. A Figura 8 mostra uma radiografia, sua segmentação e as explicações para a classificação de três arquiteturas, DenseNet-121, ResNet-18 e ResNet-152. As imagens na parte de cima são para os modelos treinados com radiografias integrais, enquanto as imagens na parte de baixo são referentes ao treinamento com radiografias segmentadas. Para cada modelo, as duas visualizações à esquerda foram obtidas com Grad-CAM, e as duas à direita com Grad-CAM++. A radiografia é de um paciente com Covid-19, e todos os três modelos classificaram a imagem corretamente.

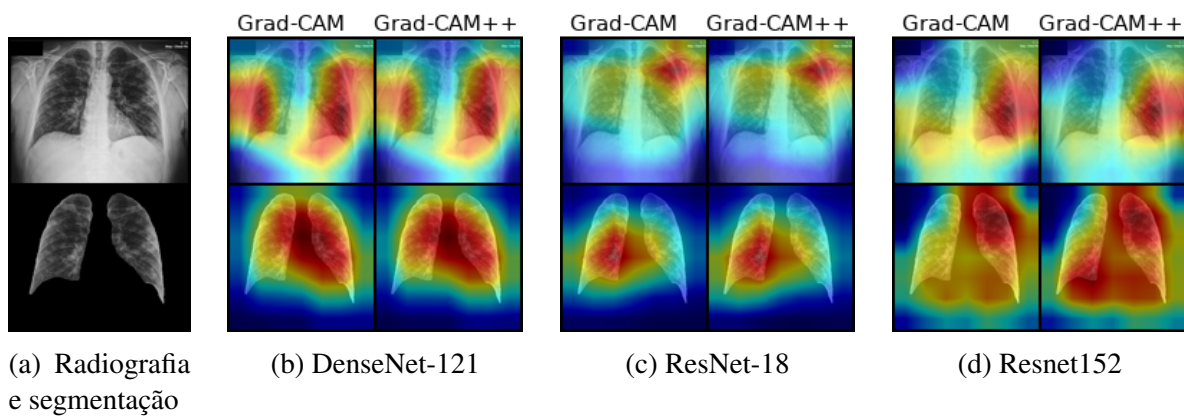


Figura 8: Visualizações geradas por Grad-CAM(++) de radiografia da classe Covid-19 para três arquiteturas

Na maioria das imagens, as duas técnicas produziram resultados semelhantes.

5.2.1 Avaliação empírica das explicações

A possibilidade de vazamento de dados e *overfitting* foi explorada na Introdução. Em suma, a mistura de datasets de diferentes fontes e a alta capacidade de *overfitting* das redes convolucionais pode levar a resultados aparentemente bons, mas que não necessariamente são reproduzíveis em larga escala. Durante a avaliação das métricas dos classificadores, a redução no desempenho quando os modelos foram treinados nas imagens segmentadas contribuiu para a hipótese de vazamento de dados.

Assim, as explicações geradas foram inspecionadas para conseguir um entendimento dos elementos nas radiografias importantes para os modelos.

Naturalmente, é de se esperar que as explicações destaquem regiões das radiografias contidas dentro dos pulmões, pois isso indicaria um entendimento pelos classificadores de que as características definidoras das classes estão nos pulmões. A Figura 9 exhibe

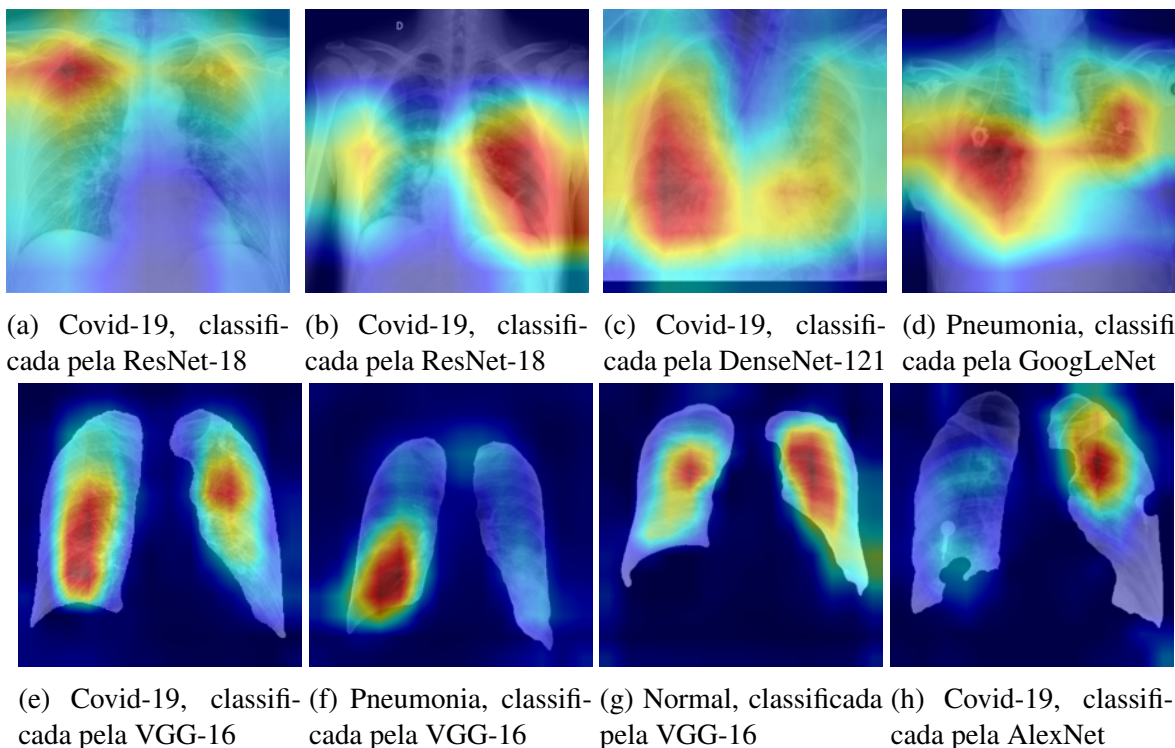


Figura 9: Explicações geradas por Grad-CAM destacando regiões dentro dos pulmões

instâncias para as quais as explicações fornecidas pelo Grad-CAM estão concentradas dentro do pulmão.

A indicação de *overfitting* ou vazamento de dados pode ser descoberta com o destaque nas explicações geradas de regiões alheias ao pulmão e, presumivelmente, ao problema geral de detecção de doenças pulmonares. Na Figura 10, são exibidas diversas instâncias em que regiões do corpo fora do pulmão estão destacadas. De maneira geral, o destaque fora do pulmão foi muito comum entre modelos treinados com as radiografias integrais, e foi observado que o treinamento com radiografias segmentadas reduziu esse efeito.

Na Figura 10, os itens (a) e (b) mostram que a clavícula e o pescoço foram regiões de interesse para a classificação. As figuras (c) e (d) mostram grande destaque para a região inferior entre os pulmões. Os outros itens mostram destaques para regiões vazias em radiografias segmentadas, geralmente entre os pulmões.

Finalmente, a Figura 11 mostra radiografias em que artefatos introduzidos de maneira intencional na imagem foram utilizados para realizar a classificação.

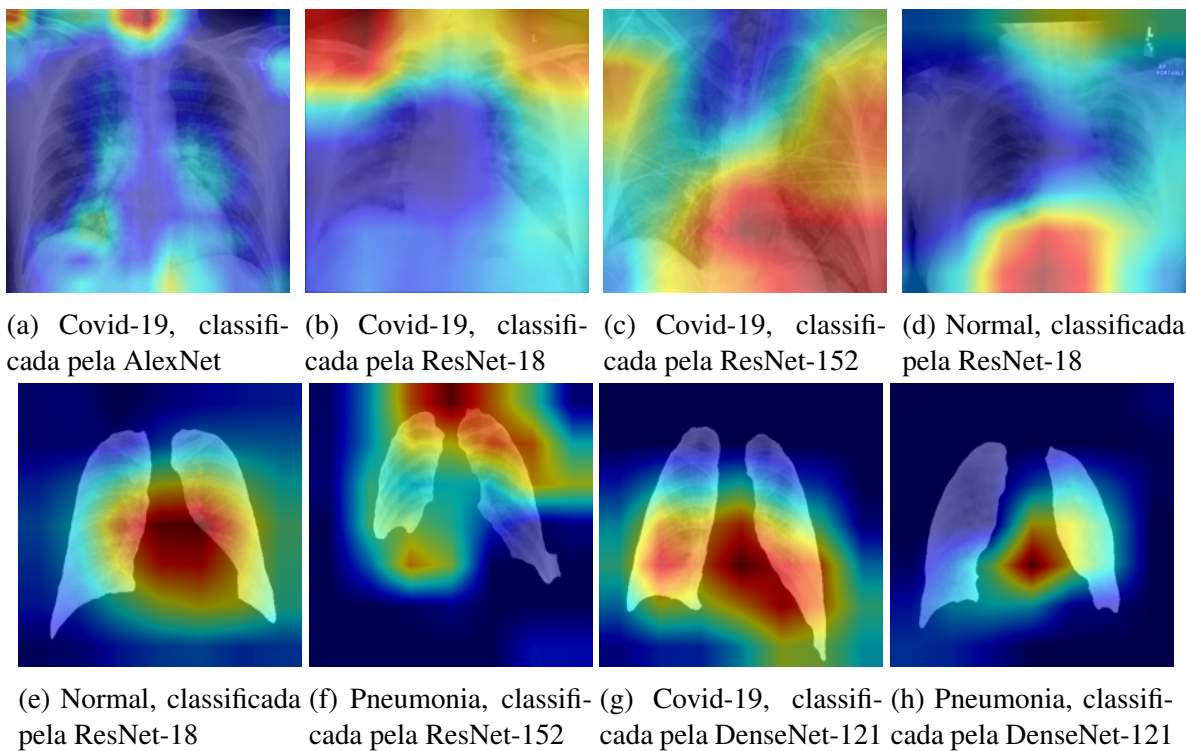


Figura 10: Explicações geradas por Grad-CAM destacando regiões fora dos pulmões

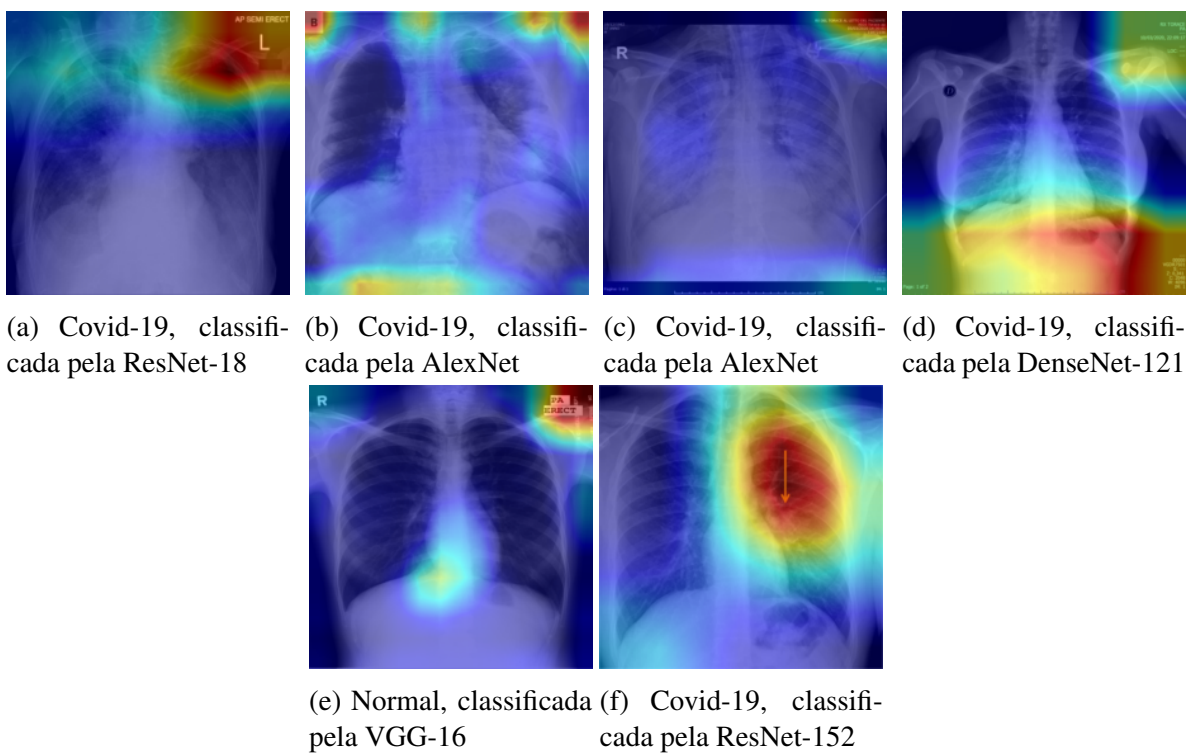


Figura 11: Explicações geradas por Grad-CAM destacando

5.2.2 Comparação das explicações com observações radiológicas

O *COVID-19 Chest X-ray Segmentations Dataset* disponibiliza anotações de achados radiológicos para algumas radiografias de pacientes com Covid-19, com segmentações para regiões de lesões pulmonares.

Comparando as segmentações disponibilizadas, é possível avaliar as explicações fornecidas pelas técnicas de interpretabilidade e, por conseguinte, é possível avaliar a confiabilidade dos classificadores, observando se as anotações de lesões coincidem com os destaques nas explicações.

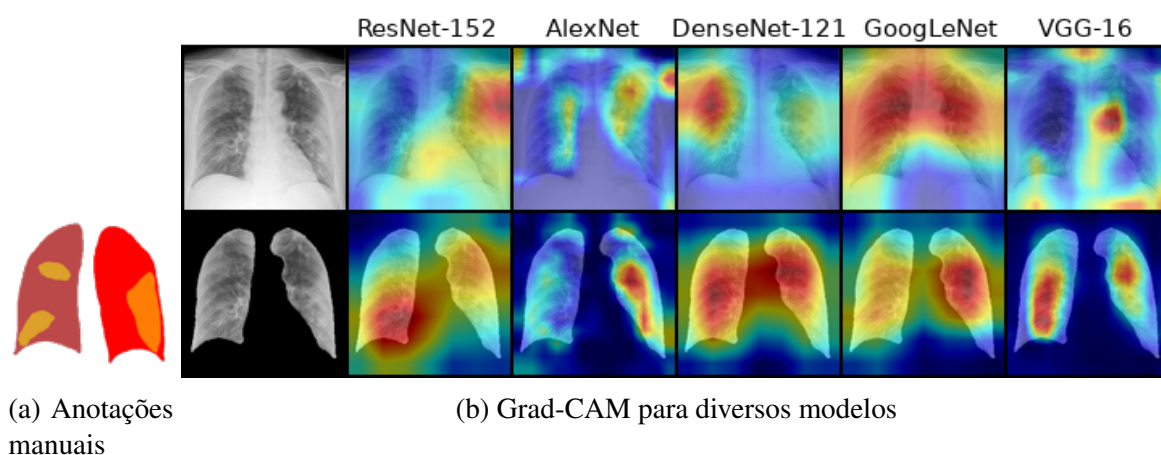


Figura 12: Comparação entre anotação manual e explicações automáticas de uma radiografia

A Figura 12 exibe essa comparação para uma radiografia. A Figura 12a destaca os achados radiológicos e a Figura 12b exibe as explicações de cinco modelos para a classificação dessa radiografia. As imagens na linha superior são para modelos treinados com as imagens integrais, enquanto a linha inferior para os modelos treinados com imagens segmentadas. Pode-se observar que, de maneira geral, as explicações não concordam com as anotações manuais. Apenas três explicações concordam minimamente, a saber: AlexNet (radiografia segmentada) e VGG-16 (ambas imagens).

A Figura 13, arranjada da mesma forma que a Figura 12, exibe outra comparação. Novamente, nota-se a falta de semelhança entre a anotação manual e as explicações geradas. Em particular, as explicações para as classificações da radiografia integral por várias modelos destacam elementos textuais na imagem.

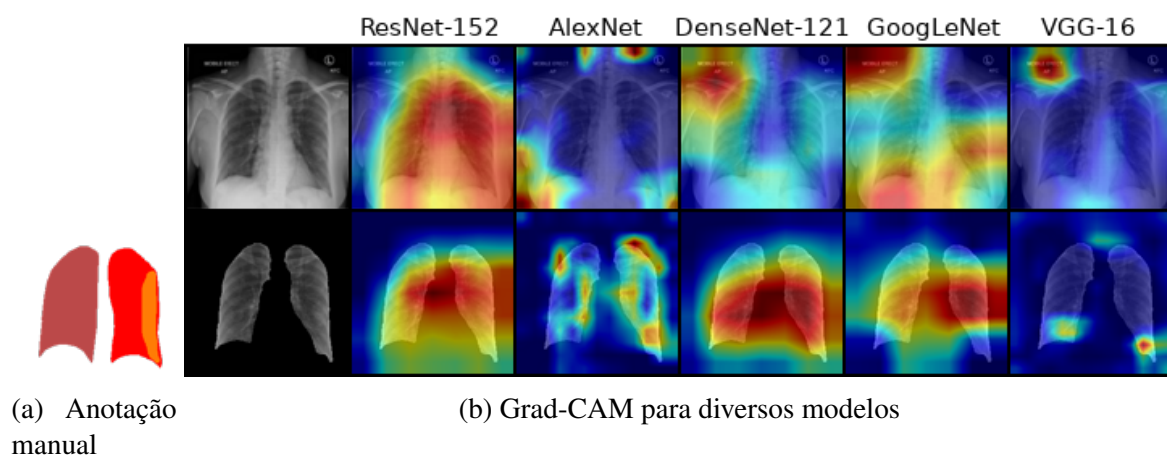


Figura 13: Comparação entre anotação manual e explicações automáticas de uma radiografia

Esses achados se generalizam para outras radiografias.

5.3 Considerações finais

Essa seção apresentou as métricas utilizadas na avaliação dos modelos treinados e detalhou os resultados obtidos. Em seguida, foi realizada uma análise das explicações para as classificações obtidas com Grad-CAM e Grad-CAM++, juntamente com uma comparação destas com anotações manuais de lesões pulmonares. Os resultados são sumariados na próxima seção.

6 Conclusão

Os modelos experimentados alcançaram ótimos resultados, especialmente os modelos treinados com radiografias integrais. Contudo, quando as mesmas arquiteturas foram treinadas com radiografias segmentadas, o desempenho foi afetado. Isso sugeriu que os modelos obtêm informações relevantes das áreas das radiografias que estão fora dos pulmões.

Após uma análise das explicações para as classificações obtidas com Grad-CAM e Grad-CAM++, foram apresentadas instâncias em que os modelos examinam os pulmões para fazer uma decisão, bem como diversas instâncias em que regiões fora do pulmão influenciaram na classificação. Em particular, foi mostrado que elementos textuais podem constituir características importantes para a classificação. Uma comparação das explicações com anotações manuais de achados radiológicos também falhou em encontrar uma semelhança entre as explicações geradas para as classificações e as lesões anotadas manualmente.

Conclui-se, portanto, que a combinação dos datasets disponíveis para treinamento e as arquiteturas utilizadas para a classificação não estão ainda confiáveis para ser utilizadas em produção, sendo necessários trabalhos futuros para determinar sua confiabilidade ou que proponham melhorias que levem a essa confiança.

7 Referências

- [1] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, “A novel coronavirus outbreak of global health concern,” *The Lancet*, vol. 395, pp. 470–473, Feb. 2020.
- [2] World Health Organization, “WHO Director-General’s opening remarks at the media briefing on COVID-19 - 11 March 2020.” <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>.
- [3] World Health Organization, “COVID-19 Weekly Epidemiological Update.” <https://www.who.int/docs/default-source/coronaviruse/situation-reports/weekly-epi-update-14.pdf>.
- [4] M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, and L. F. P. Ng, “The trinity of COVID-19: immunity, inflammation and intervention,” *Nature Reviews Immunology*, vol. 20, pp. 363–374, Apr. 2020.
- [5] A. J. Rodriguez-Morales, J. A. Cardona-Ospina, E. Gutiérrez-Ocampo, R. Villamizar-Peña, Y. Holguin-Rivera, J. P. Escalera-Antezana, L. E. Alvarado-Arnez, D. K. Bonilla-Aldana, C. Franco-Paredes, A. F. Henao-Martinez, *et al.*, “Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis,” *Travel medicine and infectious disease*, p. 101623, 2020.
- [6] D. Kermany, M. Goldbaum, W. Cai, C. Valentim, H.-Y. Liang, S. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, and K. Zhang, “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning,” *Cell*, vol. 172, pp. 1122–1131.e9, 02 2018.
- [7] A. K. Jilani, “Identifying Leakage in Computer Vision on Medical Images.” <https://www.datarobot.com/blog/identifying-leakage-in-computer-vision-on-medical-images/>.
- [8] A. LeNail, “NN-SVG: Publication-Ready Neural Network Architecture Schematics,” *Journal of Open Source Software*, vol. 4, no. 33, p. 747, 2019.
- [9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [10] W. Rawat and Z. Wang, “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review,” *Neural Computation*, vol. 29, pp. 2352–2449, Sept. 2017.

- [11] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation Applied to Handwritten Zip Code Recognition,” *Neural Computation*, vol. 1, pp. 541–551, Dec. 1989.
- [12] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson, *Handwritten Digit Recognition with a Back-Propagation Network*, pp. 396–404. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990.
- [13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, May 2015.
- [14] K. Chellapilla, S. Puri, and P. Simard, “High Performance Convolutional Neural Networks for Document Processing,” in *Tenth International Workshop on Frontiers in Handwriting Recognition* (G. Lorette, ed.), (La Baule (France)), Université de Rennes 1, Suvisoft, Oct. 2006. <http://www.suvisoft.com>.
- [15] F. Nasse, C. Thureau, and G. A. Fink, “Face Detection Using GPU-Based Convolutional Neural Networks,” in *Computer Analysis of Images and Patterns*, pp. 83–90, Springer Berlin Heidelberg, 2009.
- [16] S. C. Turaga, J. F. Murray, V. Jain, F. Roth, M. Helmstaedter, K. Briggman, W. Denk, and H. S. Seung, “Convolutional Networks Can Learn to Generate Affinity Graphs for Image Segmentation,” *Neural Computation*, vol. 22, pp. 511–538, Feb. 2010.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, pp. 84–90, May 2017.
- [18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, pp. 211–252, Apr. 2015.
- [19] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun, “What is the best multi-stage architecture for object recognition?,” in *2009 IEEE 12th International Conference on Computer Vision*, IEEE, Sept. 2009.
- [20] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML’10*, (Madison, WI, USA), pp. 807–814, Omnipress, 2010.

- [21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0580, 2012.
- [22] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [25] S. Li, Z.-Q. Liu, and A. B. Chan, “Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 482–489, 2014.
- [26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.
- [27] S. S. Farfade, M. J. Saberian, and L.-J. Li, “Multi-View Face Detection Using Deep Convolutional Neural Networks,” in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ICMR ’15*, (New York, NY, USA), pp. 643–650, Association for Computing Machinery, 2015.
- [28] M. Liang, W. Tang, D. M. Xu, A. C. Jirapatnakul, A. P. Reeves, C. I. Henschke, and D. Yankelevitz, “Low-Dose CT Screening for Lung Cancer: Computer-aided Detection of Missed Lung Cancers,” *Radiology*, vol. 281, no. 1, pp. 279–288, 2016. PMID: 27019363.
- [29] T. Kooi, G. Litjens, B. Ginneken, A. Gubern-Mérida, C. Sánchez, R. Mann, G. Heeten, and N. Karssemeijer, “Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions,” *Medical Image Analysis*, vol. 35, 08 2016.
- [30] V. Chouhan, S. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damasevicius, and V. Albuquerque, “A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images,” *Applied Sciences*, vol. 10, p. 559, 01 2020.

- [31] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.
- [32] 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,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017.
- [33] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, IEEE, 2018.
- [34] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?” Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- [35] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: Automatic detection from X-Ray images utilizing Transfer Learning with Convolutional Neural Networks,” *Physical and Engineering Sciences in Medicine*, p. 1, 2020.
- [36] S. H. Kassani, P. H. Kassasni, M. J. Wesolowski, K. A. Schneider, and R. Deters, “Automatic Detection of Coronavirus Disease (COVID-19) in X-Ray and CT Images: A Machine Learning-Based Approach,” *arXiv preprint arXiv:2004.10641*, 2020.
- [37] Y. Oh, S. Park, and J. C. Ye, “Deep Learning COVID-19 Features on CXR using Limited Training Data Sets,” *IEEE Transactions on Medical Imaging*, 2020.
- [38] N. Punj and S. Agarwal, “Automated diagnosis of COVID-19 with limited posteroanterior chest X-Ray images using fine-tuned deep neural networks,” *Applied Intelligence*, pp. 1–14, 10 2020.
- [39] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, “COVID-19 Screening on Chest X-ray Images Using Deep Learning based Anomaly Detection,” *arXiv preprint arXiv:2003.12338*, 2020.
- [40] B. Ghoshal and A. Tucker, “Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection,” *arXiv preprint arXiv:2003.10769*, 2020.

- [41] T. Li, Z. Han, B. Wei, Y. Zheng, Y. Hong, and J. Cong, “Robust Screening of COVID-19 from Chest X-ray via Discriminative Cost-Sensitive Learning,” *arXiv preprint arXiv:2004.12592*, 2020.
- [42] E. Luz, P. L. Silva, R. Silva, and G. Moreira, “Towards an Effective and Efficient Deep Learning Model for COVID-19 Patterns Detection in X-ray Images,” *arXiv preprint arXiv:2004.05717*, 2020.
- [43] L. Wang, Z. Q. Lin, and A. Wong, “COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images,” *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.
- [44] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, “Automated detection of COVID-19 cases using deep neural networks with X-ray images,” *Computers in Biology and Medicine*, p. 103792, 2020.
- [45] H. S. Maghdid, A. T. Asaad, K. Z. Ghafoor, A. S. Sadiq, and M. K. Khan, “Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms,” *arXiv preprint arXiv:2004.00038*, 2020.
- [46] J. Born, G. Brändle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, and N. Wiedemann, “POCOVID-Net: automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS),” *arXiv preprint arXiv:2004.12084*, 2020.
- [47] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017.
- [48] D. Kermany, “Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images,” 2018.
- [49] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.
- [50] B. Van Ginneken, M. B. Stegmann, and M. Loog, “Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database,” *Medical image analysis*, vol. 10, no. 1, pp. 19–40, 2006.

- [51] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases,” *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.
- [52] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haggoo, R. Ball, K. Shpanskaya, *et al.*, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597, 2019.
- [53] J. P. Cohen, P. Morrison, and L. Dao, “COVID-19 image data collection,” *arXiv 2003.11597*, 2020.
- [54] “CoronaHack -Chest X-Ray-Dataset | Kaggle.” <https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset>.
- [55] “COVID-19 X rays | Kaggle.” <https://www.kaggle.com/andrewmvd/convid19-X-rays>.
- [56] R. Selvan, E. Dam, N. Detlefsen, S. Rischel, K. Sheng, M. Nielsen, and A. Pai, “Lung Segmentation from Chest X-rays using Variational Data Imputation,” 07 2020.