UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LUIZA CARVALHO SILVEIRA

COMBINAÇÃO DE TÉCNICAS DE APRENDIZADO PROFUNDO PARA A CLASSIFICAÇÃO DE RAIO-X TORÁCICO EM APOIO AO DIAGNÓSTICO DE COVID-19

Recife

2022

LUIZA CARVALHO SILVEIRA

COMBINAÇÃO DE TÉCNICAS DE APRENDIZADO PROFUNDO PARA A CLASSIFICAÇÃO DE RAIO-X TORÁCICO EM APOIO AO DIAGNÓSTICO DE COVID-19

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Inteligência Computacional.

Orientador: Dr. Fernando Maciano de Paula Neto

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S587c Silveira, Luiza Carvalho
Combinação de técnicas de aprendizado profundo para a classificação de raio-x torácico em apoio ao diagnóstico de Covid-19 / Luiza Carvalho Silveira. – 2022.
228 f.:il., fig, tab.

Orientador: Fernando Maciano de Paula Neto.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.

Inclui referências.

1. Inteligência computacional. 2. Aprendizado profundo. I. Paula Neto, Fernando Maciano de (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2022-193

LUIZA CARVALHO SILVEIRA

“COMBINAÇÃO DE TÉCNICAS DE APRENDIZADO PROFUNDO PARA A CLASSIFICAÇÃO DE RAIO-X TORÁCICO EM APOIO AO DIAGNÓSTICO DE COVID-19”

Dissertação apresentada ao Programa de Pós Graduação em Ciências da Computação da Universidade Federal de Pernambuco, Centro Acadêmico de informática, como requisito para a obtenção do título de Mestre em Ciência da computação. Área de concentração: Inteligência Computacional.
.

Aprovado em: 25/08/2022.

BANCA EXAMINADORA

_____
Prof. Dr. Fernando Maciano de Paula Neto **(Orientador)**
Universidade Federal de Pernambuco - UFPE

_____
Prof. Dr. Aluízio Fausto Ribeiro Araújo
Centro de Informática / UFPE

_____
Profª. Drª. Thais Gaudencio Do Rego
DI - CI - Departamento de Informática / UFPB

Dedico este trabalho a minha família e minha namorada que foram porto seguro perante as dificuldades durante este percurso.

AGRADECIMENTOS

Aos meus pais, Antônio e Micheline e ao meu irmão Vinicius, pela confiança no meu progresso, cobranças e pelo apoio emocional. A Aninha, por ter sido sempre presente me apoiando em tudo e uma pessoa exemplar, a qual eu sempre serei grata por ter convivido.

A minha namorada, Ana Flávia, cujo apoio, revisões de texto, incentivo e companheirismo foram essenciais para que fosse possível finalizar meu mestrado. Aos meus amigos, Nícolas, Deborah, Ramon e Williany por aguentarem minha indisponibilidade constante e me apoiarem nas minhas dificuldades.

Ao meu orientador, Fernando Maciano, que sempre se mostrou muito compreensivo e solícito. As suas valiosas indicações e mentorias fizeram toda a diferença.

A todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho.

“Eu acredito que às vezes, são as pessoas de quem ninguém espera nada que fazem as coisas que ninguém consegue imaginar” (HODGES, 1983).

RESUMO

A aplicação de técnicas de *deep learning* no âmbito de serviços de saúde é um campo de pesquisa emergente na área de Ciência da Computação. A pandemia da COVID-19 motivou o desenvolvimento de modelos de *deep learning* para a detecção de padrões de imagem para a utilização alternativa desses exames detecção da síndrome respiratória. Este trabalho tem como objetivo a realização de uma análise exploratória combinando diferentes técnicas anteriormente aplicadas para detecção da COVID-19, o *transfer learning*, o aumento de dado pela geração de imagens sintéticas por meio de GANs e a segmentação de imagens, de modo a obter diferentes *pipelines* e descrever a qualidade das melhores encontradas para diferenciação de três classes de pacientes: aqueles infectados pelo vírus da COVID-19, pacientes com síndrome respiratória causada por outro agente que não o SARS-CoV-2, e, finalmente, pacientes saudáveis. Além disso, é aberta uma discussão sobre a real capacidade de generalização dos modelos existentes até então ao se realizar predições para conjuntos de dados de teste secundários e nunca vistos pelo modelo, assim testando a capacidade do mesmo de predizer resultados para dados com características diferentes aos do seu conjunto de treino. Foram criadas quatro *pipelines* diferentes para dois conjuntos de dados, com a *pipeline* proposta obtendo um *F1-Score* de 90,8% para o conjunto de testes e 55,1% para o conjunto de testes secundário.

Palavras-chave: aprendizado profundo; COVID-19; classificação multiclasse; segmentação de imagens; raio-x torácico; pneumonia.

ABSTRACT

The application of deep learning techniques in health services is an emerging field of research in Computer Science. The COVID-19 pandemic motivated the development of deep learning models for the detection of image patterns for the alternative use of these exams detection of the respiratory syndrome. This paper aims to perform an exploratory analysis combining different techniques previously applied for COVID-19 detection, such as transfer learning, data augmentation by generating synthetic images through GANs and image segmentation, in order to obtain different pipelines and describe the quality of the best ones found for differentiating three classes of patients: those infected with the COVID-19 virus, patients with respiratory syndrome caused by an agent other than SARS-CoV-2, and finally healthy patients. In addition, a discussion is opened about the real generalizability of the existing models so far when making predictions for secondary test data sets never seen by the model, thus testing it's ability to predict results for data with different characteristics than its training set. Four different pipelines were created for two data sets, with the proposed pipeline obtaining an F1-Score of 90.8% for the test set and 55.1% for the secondary test set.

Keywords: deep learning; COVID-19; multiclass classification; image segmentation; chest x-ray; pneumonia.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

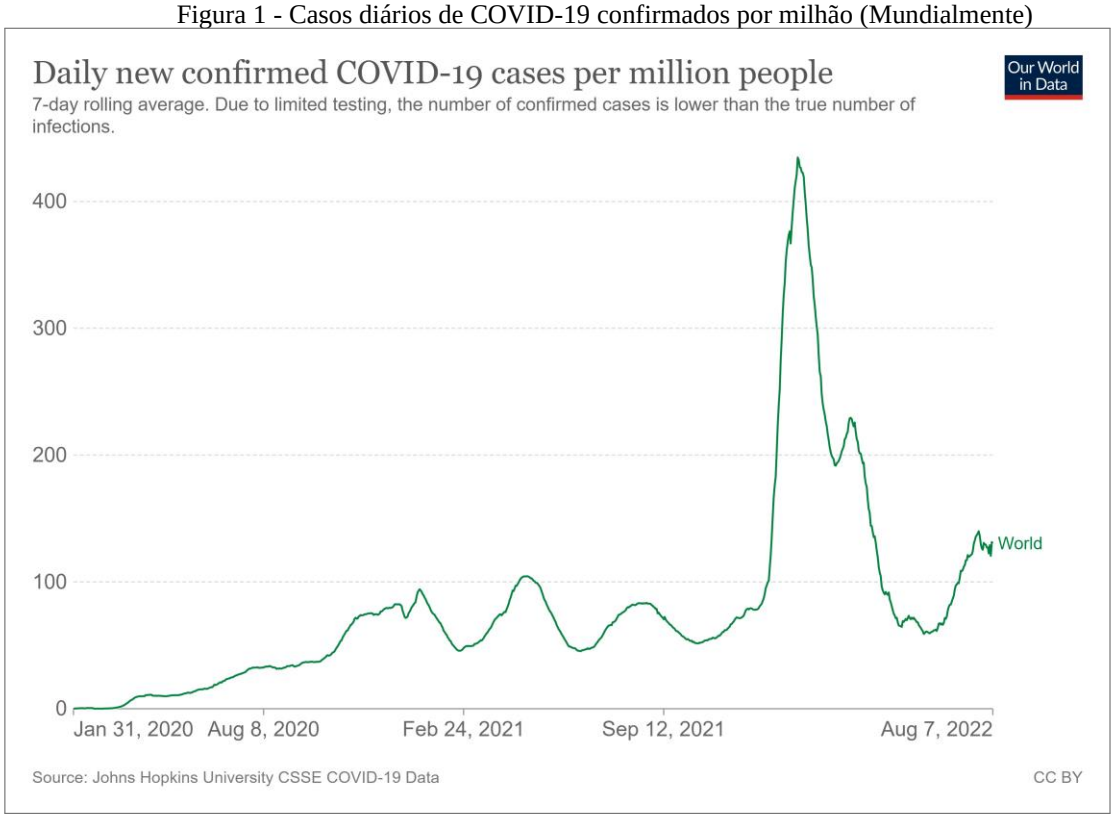
ADAS	*Advanced Driver Assistance*
ANNs	*Artificial Neural Networks*
AUC	*Area Under The Curve*
BRISK	*Binary Robust Invariant Scalable Keypoint*
CNNs	*Convolutional Neural Networks*
DA	*Data Augmentation*
DCGANs	*Deep Convolutional Generative Adversarial Networks*
DNNs	*Deep Neural Networks*
FGV	Fundação Getúlio Vargas
FN	*False Negatives*
FP	*False Positives*
FPR	*False Positives Rate*
FGV	Fundação Getúlio Vargas
GANs	*Generative Adversarial Networks*
GPU	*Graphic Processing Units*
IA	Inteligência Artificial
MATLAB	*MATrix LABoratory*
MNIST	*Modified National Institute of Standards and Technology Large*
OCT	*X-Ray Images*
OMS	Organização Mundial da Saúde
ReLu	*Rectified Linear Unit*
ROC	*Receptor Operator Characteristic*
RX	Raio-X
SARS	*Severe Acute Respiratory Syndrome*
SNNs	*Simulated Neural Networks*
TC	Tomografia
TP	*True Positive*
TPR	*True Positives Rate*

SUMÁRIO

1 INTRODUÇÃO

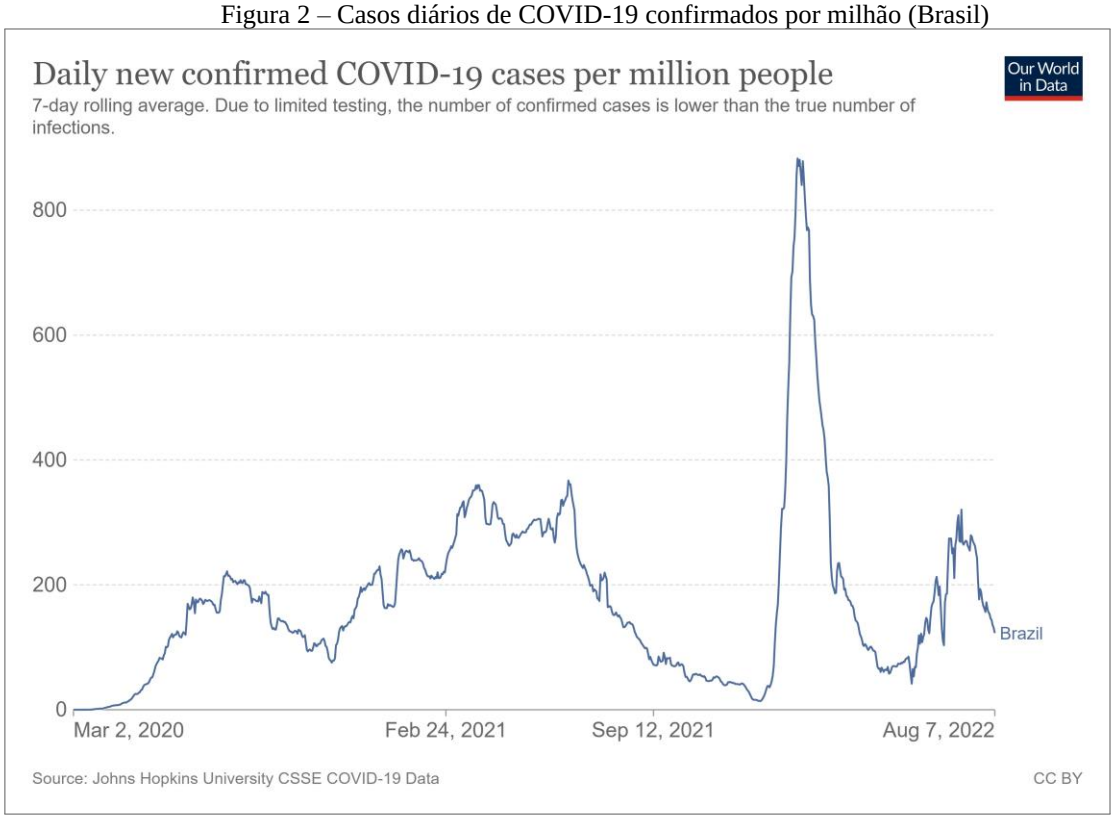
A Organização Mundial da Saúde (OMS) declarou em março de 2020 a disseminação de casos de COVID-19 como situação pandêmica (WHO, 2020), sendo a primeira pandemia desencadeada por um coronavírus que, no caso da COVID-19, é o SARS-CoV-2. Os coronavírus são associados à síndrome respiratória aguda grave, ou *severe acute respiratory syndrome* (SARS), um tipo de doença respiratória viral (KHAN et al., 2020). Desde os primeiros casos de infecção de COVID-19 registrados em dezembro de 2019 em Wuhan, província de Hubei, na China (LI et al., 2020), a OMS identificou variantes do novo coronavírus SARS-CoV-2 como por exemplo a Ômicron, associada como a mais infecciosa até então e, por isso, sendo o agente viral causador do maior pico de casos observados até o momento. De acordo com dados da OMS (OMS, 2022) até agosto de 2022, globalmente, já foram notificados e confirmados em torno de 585 milhões de casos de COVID-19, dos quais, no total, foram registrados em torno de 6,42 milhões de mortos. No Brasil, segundo o Ministério da Saúde (SAUDE, 2022), já foram registrados cerca de 34 milhões de casos e um total de aproximadamente 680 mil mortes. Os dados expostos podem ser observados nas Figuras 1 e 2.

Figura 1 - Casos diários de COVID-19 confirmados por milhão (Mundialmente)



Fonte: (MATHIEU EDOUARD; RITCHIE, 2021)

Figura 2 – Casos diários de COVID-19 confirmados por milhão (Brasil)



Fonte: (MATHIEU EDOUARD; RITCHIE, 2021)

Tendo em vista os grandes impactos causados pelo novo coronavírus, tais como sobrecarga dos sistemas de saúde, repercussões negativas na economia mundial, e até efeitos indiretos de perdas de aprendizagem entre crianças que estão fora da escola, segundo a Fundação Getúlio Vargas (FGV) em 2019, a taxa de crianças fora das escolas era de 1,39%. Em 2020, esse número saltou para 5,5% (NERI; OSORIO, 2022).

A comunidade científica configura-se, desde o começo da pandemia, como uma importante fonte de apoio a decisões relacionadas à segurança sanitária e de saúde, o que envolve a promoção de pesquisa e produção de informação estratégica para ampliar o debate sobre o tema e orientar a tomada de decisão, considerando os desafios enfrentados pelos governos e nações, individual e coletivamente, para a execução de medidas em resposta à pandemia. Devido ao alto nível de transmissibilidade da COVID-19, uma das principais medidas que podem ser adotadas com o objetivo de controlar sua disseminação e recuperar os pacientes é melhorar a identificação e rastreio de infecções, para então introduzir o tratamento imediato e isolamento delas (PUNN; AGARWAL, 2020).

Testes laboratoriais de detecção viral do tipo RT-PCR ou de detecção de anticorpos, do tipo sorológicos, até o momento de publicação deste trabalho, são os principais métodos de detecção da COVID-19. Entretanto, uma limitação comum em testes laboratoriais é o tempo de espera

até a obtenção de resultado, o que, por sua vez, pode atrasar a adoção do tratamento adequado em tempo hábil. Como consequência, a doença pode se agravar e contribuir para manifestação mais grave das sequelas ou progressão à morte. Além disso, os exames de sangue e RT-PCR comportam janelas específicas de tempo em que são eficazes, podendo resultar em falsos negativos se não realizados na janela incorreta de tempo (OSMAN; DAAJANI; ALSAHAFI, 2020). Também há problemas quanto às restrições de acesso aos testes laboratoriais e a sua escassez, ocasionando dificuldades na testagem em massa, especialmente para a população de baixa renda e nas redes de saúde pública de países como o Brasil (ROCHA, 2022).

Por outro lado, os exames por imagens podem, ainda, auxiliar na identificação de casos da síndrome pós-COVID, também conhecida como COVID longa, caracteriza pela persistência de sintomas relacionados a doença por mais de três meses. Estudos apontaram que existem mais de 200 sintomas caracterizadores da síndrome pós-COVID sendo os mais usuais: fadiga, falta de ar e disfunções cognitivas. O Centro de Controle e Prevenção de Doenças dos Estados Unidos estimou que 1 em cada 5 indivíduos pode desenvolvê-la (BULL-OTTERSON et al., 2022). Para identificação dessa nova forma da doença os testes laboratoriais não são adequados, pois estes só funcionam na fase aguda da Covid. Todavia, evidências científicas apontam para a possibilidade de diagnóstico da COVID longa por meio de exames de imagem, seja com vistas a mapear a atividade metabólica das células pulmonares via PET-CT, seja para averiguar alterações na densidade dos pulmões por meio de RX torácicos (RODRIGUES et al., 2022).

Dentre as técnicas existentes para o apoio à decisão em políticas públicas no âmbito da saúde, a aplicação de técnicas de *deep learning* em serviços de saúde emerge nas últimas décadas na área de Ciência da Computação devido à capacidade de apoio da inteligência artificial (IA), sobretudo aos profissionais médicos. Um dos campos de aplicação de modelos de *deep learning* no âmbito da medicina é no auxílio do diagnóstico de condições de saúde a partir de exames de imagem, como ocorre para a identificação de síndromes respiratórias através de tomografia de tórax (RAJPURKAR et al., 2017a), para diagnóstico de doenças cardíacas via tomografia computadorizada com emissão de fóton único (DILSIZIAN; SIEGEL, 2013), detecção de câncer de mama através de mamografia (SHEN et al., 2019), dentre outras finalidades.

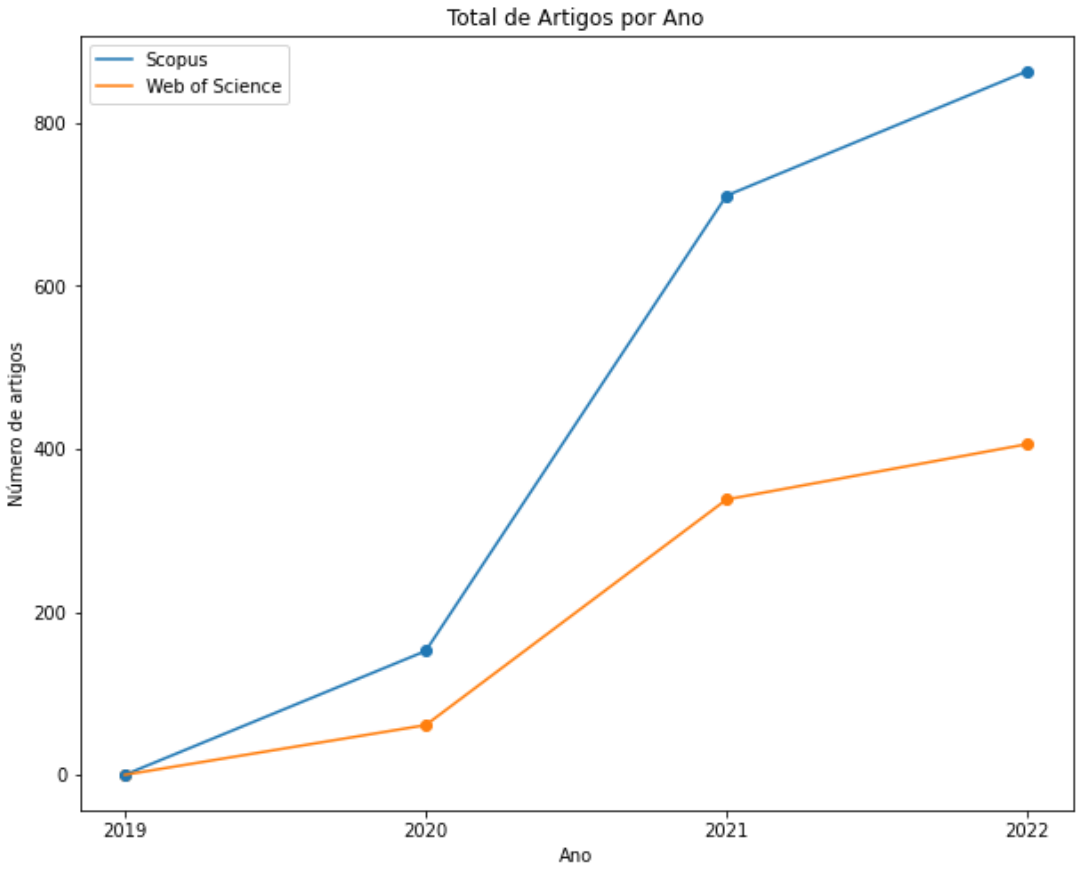
Com base na potencial contribuição de ferramentas *deep learning* ao diagnóstico via exames de imagem, desde a declaração da pandemia, observa-se um crescente interesse de pesquisadores em técnicas de aprendizado de máquina e no aprendizado profundo *(deep learning) soluções* para auxiliar no diagnóstico da COVID-19. Alguns estudos elencam exames de imagem como raio-X (RX) do tórax e tomografia (TC) como meio de detecção alternativos

aos testes de detecção RT-PCR e sorológicos (LIU et al., 2020; AI et al., 2020; BERNHEIM et al., 2020), dado que esses exames já são realizados rotineiramente em pacientes com suspeitas ou casos confirmados da doença para avaliar se há comprometimento do pulmão do infectado.

Um dos achados de estudos científicos relacionados à COVID-19 é o de que pulmões de pacientes sintomáticos mostra determinadas marcas visuais tais como opacidades de vidro fosco-escuro que podem diferenciar pacientes infectados daqueles não infectados (FANG et al., 2020; XIE et al., 2020). No estudo de Huang et al. (2020), também foi identificada a existência de anomalias radiográficas bilaterais em imagens RX na maioria dos casos positivos da COVID-19 analisados; em outro artigo (GUAN et al., 2020) foram identificados casos positivos da COVID-19 que apresentavam anomalias radiográficas, tais como opacidade de vidro fosco, anomalias bilaterais, e anomalias intersticiais em imagens tanto de RX como de TC. Apesar das imagens TC representarem um maior nível de detalhamento, e serem consideradas mais precisas (HUANG et al., 2020), o estudo de Wong et al. (2020) demonstra que achados comuns em imagens tomográficas de pacientes infectados pelo COVID-19 também podem ser identificados nos exames de raio-X de pacientes.

Dado o potencial a ser explorado na utilização de imagens de raio-X torácico para detecção da COVID-19, pesquisadores acreditam que um sistema baseado na radiologia torácica pode ser uma ferramenta eficaz na detecção, quantificação e acompanhamento dos casos COVID-19 (HUANG et al., 2020; WONG et al., 2020; WANG; LIN; WONG, 2020). Em específico, os estudos publicados têm buscado a realização da classificação de imagens radiográficas como imagens pertencentes a pacientes que apresentam pneumonia causada pelo vírus SARS-CoV-2, pneumonia causada por outro vírus ou bactéria ou pacientes sem achados, isto é, com imagens ditas como normais. Alternativamente também se tem focado apenas na detecção do COVID-19, diferenciando imagens de pacientes com diagnóstico da doença de outras imagens. Pode ser visto na Figura 3 um interesse crescente nesse tipo de pesquisa nas plataformas *Scopus* e *Web of Science* desde o início do ano de 2020:

Figura 3 - Total de artigos por ano (*Scopus e Web of Science*)



Fonte: Própria

Existem várias vantagens em utilizar imagens RX ao invés de tomografias de tórax para a detecção do COVID-19, particularmente em áreas com recursos limitados e em áreas muito afetadas, pois, além da alta disponibilidade de máquinas de raio-X em hospitais, existindo ainda dispositivos portáteis de radiografia de tórax (NG et al., 2020; RUBIN et al., 2020), é necessário também considerar também o custo e a complexidade do equipamento utilizado. Nesse contexto, além do raio-X torácico requirir um equipamento relativamente barato, ele pode ser realizado rapidamente em salas isoladas com um dispositivo portátil de radiografia de tórax, o que reduz o risco de infecção dentro dos hospitais, além de expor menos o paciente a radiação (CLAESSENS et al., 2015). Ademais, um sistema de detecção baseado na imagem radiológica do tórax tem a capacidade de analisar vários casos simultaneamente, alta disponibilidade e, sobretudo, pode ser muito útil em hospitais sem ou com um número limitado de recursos de testagem, podendo diminuir o tempo de espera por diagnóstico e facilitar o início do tratamento (WANG; LIN; WONG, 2020).

Várias soluções têm sido criadas para a detecção da COVID-19 por meio de imagens de raio-X, desde simples utilizações técnicas mais básicas de *machine learning* como os

algoritmos de *K-Nearest-Neighbors,* ou árvores de decisão para a tarefa de classificação, bem como o uso de redes neurais previamente treinadas. Há, também, a criação de redes neurais complexas com arquiteturas montadas especificamente para a solução desse problema.

Apesar da diversa gama de soluções para a detecção da COVID-19 por meio de imagens de raio-X, existem limitações a serem consideradas. Dentre elas, a utilização de um número pequeno de dados para o treinamento do modelo e a pouca variabilidade nos conjuntos de dados utilizados podem não ser suficientemente diversas para detectar, verdadeiramente, características da COVID-19, assim gerando modelos possivelmente enviesados (MAIOR et al., 2021). O uso de dados privados dificultando a reprodutibilidade, a falta de informações em relação às métricas utilizadas para avaliação dos modelos, tais como acurácia, precisão, *F1-score* e *recall* médios, também são limitações recorrentes.

Em sua maioria, os artigos publicados falam apenas do melhor resultado, e não discutem sobre médias ou desvios padrões obtidos. Também há pouca discussão sobre a capacidade de generalização dos modelos para outras bases de dados que não as utilizadas para treinamento. Apesar de os estudos publicados até o momento de publicação deste trabalho executarem bem a classificação de imagens, o que pode ser afirmado a partir das altas taxas de acurácia reportadas na literatura, esse fenômeno pode não ser refletido na performance do modelo no meio hospitalar, porque um modelo pode facilmente se tornar especialista em sua base de dados de treino e não há como se saber ao certo quais características ele leva em consideração ao realizar a classificação. Dessa forma, o modelo se torna capaz de aprender características específicas das bases de dados ao qual está sendo treinado e, assim, não consegue obter um bom desempenho para imagens nunca vistas antes e que apresentam características diferentes.

Alguns problemas existentes nos métodos de detecção de COVID-19 por imagens radiográficas ganharam atenção e pesquisadores tentam solucioná-los, por exemplo, a baixa disponibilidade de dados vem sendo contornada com a utilização de técnicas de aumento de dados ou *data augmentation* (DA), que podem variar desde técnicas convencionais como as transformações geométricas, aumento de brilho ou contraste, corte, rotação, injecção de ruído, e outros, até a geração de imagens sintéticas com o auxílio de *generative adversarial networks* (GANs) (LOEY; SMARANDACHE; KHALIFA, 2020; UMER et al., 2021). Existem também trabalhos nos quais se tenta diminuir o enviesamento dos modelos utilizando bases de dados com maior variabilidade ou até realizando uma segmentação das imagens de raio-X para que a rede analise apenas os pulmões dos pacientes (MAIOR et al., 2021; TEIXEIRA et al., 2021).

Desse modo, este trabalho tem como objetivo explorar combinações de técnicas já vistas na literatura e buscar uma possível combinação ótima para executar a tarefa da classificação de

imagens radiográficas com o objetivo de diferenciar pacientes que apresentam pneumonia causada pelo vírus SARS-CoV-2, pneumonia causada por outro vírus ou bactéria e pacientes sem achados, ditos normais. Também serão utilizados em conjunto diferentes combinações de bases de dados abertos com o objetivo de aumentar a robustez e testar a capacidade de generalização de cada *pipeline* criado. A classificação multiclasse foi escolhida em detrimento da binária pelo fato de que muitos pacientes com pneumonia viral ou bacteriana podem ter sintomas similares aos da COVID-19, e em um sistema de classificação binário esses pacientes poderiam acabar sendo agrupados erroneamente como pacientes com esse vírus ou serem agrupados em conjunto com pacientes sem achados, assim sendo prejudicados em seu tratamento.

O presente estudo une três técnicas com intuito de diminuir o aumento artificial dos resultados detectado por Maior et al. (2021) e, assim, oferecer uma discussão sobre a possíveis vantagens de utilizar combinações de métodos variados e vários conjuntos de dados abertos para o treinamento de um modelo de inteligência artificial. Como objetivo secundário, analisa-se, para a problemática proposta, se a aplicação de aumento de dados é algo benéfico para ao modelo; ainda, são analisadas técnicas de segmentação de imagens com o objetivo de extrair características realmente relevantes à classificação almejada. Enfim, com diferentes combinações dessas abordagens, se espera evitar um enviesamento do modelo, de modo que ele acabe detectando características dos pulmões apresentados pelas imagens RX, e não características específicas de cada fonte de dados.

Também é proposta uma análise sobre a capacidade de generalização dos modelos criados a partir das *pipelines* propostas. Essa análise é feita por meio da utilização de um conjunto de testes secundário, no qual são utilizadas 2 bases de dados nunca vistas pelos modelos para verificar se as características das imagens captadas por ele são úteis para identificar as classes em conjuntos de imagens provenientes de outra fonte.

Mais especificamente foram desenvolvidos 4 *pipelines* combinando as diferentes técnicas citadas com diferentes combinações de 8 bases de dados abertas e 3 modelos já existentes de redes neurais convolucionais (CNNs) para distinguir imagens radiológicas de pacientes sem achados e infectados, diferenciando estes últimos entre pacientes com pneumonia causada pelo COVID-19 e pneumonia causada por outro vírus ou bactéria. Posteriormente, foi comparada a performance de teste dos *pipelines* criados e sua performance em relação a uma base que nunca foi vista pelos modelos envolvidos no processo, chamada de base de testes secundária.

O restante deste estudo se estrutura da seguinte forma: ainda neste capítulo, é realizado um apanhado sobre os trabalhos relacionados. No segundo capítulo, revisando a literatura brevemente, apresenta-se uma visão geral sobre aprendizagem de máquina, aprendizado profundo e CNNs, e, também, uma revisão bibliográfica dos modelos que foram utilizados nas construções de *pipeline* deste trabalho. O terceiro capítulo detalha as bases de dados de imagens RX utilizadas, assim como a metodologia proposta. Os resultados obtidos são apresentados e discutidos no quarto capítulo. Finalmente, o quinto capítulo resume as principais conclusões do trabalho e apresenta suas limitações e observações finais.

1.1 TRABALHOS RELACIONADOS

Na presente seção, são discutidos artigos relacionados a algum dos seguintes tópicos: diagnóstico da COVID-19 em imagens RX, segmentação de imagens de RX pulmonares e utilização de GANs como forma de aumento de dados para classificação de COVID-19. Também são discutidos potenciais limitações, enviesamentos e problemas de identificação da COVID-19, dado o estado atual das bases de dados disponíveis. O quadro 1 apresenta os artigos analisados nesta seção, juntamente com suas acurácias e *F1-Score* de teste e metodologias aplicadas. Todos os artigos analisados tratam da classificação multiclasse com 3 classes, COVID-19, pneumonia e pulmão normal. O quadro 2 apresenta detalhamentos sobre as bases de dados utilizadas pelos artigos aqui citados.

É importante observar que, devido ao agravamento da pandemia, a identificação da COVID-19 em imagens de RX tornou-se um tema muito debatido, sendo inviável representar o atual estado da arte para essa tarefa, uma vez que novas obras estão sendo publicadas todos os dias. Portanto, foram escolhidos artigos com relevância e diretamente relacionados ao presente estudo para serem debatidos.

Dentre os artigos elencados, o trabalho de Wang, Lin e Wong (2020) se destaca. Os autores propuseram uma CNN chamada COVID-Net para a classificação de imagens de RX torácico usando dados abertos disponíveis ao público geral. Uma grande contribuição deste trabalho é a introdução do COVIDx, um conjunto de dados de referência de acesso aberto compreendendo 13.975 imagens RX em 13.870 casos de pacientes. Com uma acurácia de 93.3%, a COVID-Net é uma das soluções mais robustas da literatura até então, por utilizar um conjunto de dados grande e variado, uma arquitetura própria voltada para a detecção da COVID capaz de explicar os resultados da rede proposta utilizando *GSInquire*. No entanto, não são disponibilizados dados experimentais sobre as médias e desvios de padrões das métricas

utilizadas, o que faz com que se assuma que o artigo apresenta apenas a melhor performance do modelo criado.

MAIOR et al. também apresenta uma rede de arquitetura própria como solução para o problema de classificação da COVID-19. O banco de dados utilizado também é o COVIDx, mas é adicionada outra base de dados ao conjunto (*Large Dataset of Labeled Optical Coherence Tomography and Chest X-Ray Images* (OCT)) e aplicado o aumento de dados apenas na classe de pacientes infectados pela COVID-19, utilizando rotações, espelhamento e adição de ruído. O artigo tenta buscar todas as classes na maioria dos conjuntos utilizados, para tentar diminuir, desse modo, o enviesamento do modelo, evitando que ele aprenda apenas a distinguir características específicas das bases de dados utilizadas.

É feita uma crítica por parte dos autores às limitações em relação à falta de capacidade de processamento de imagens com alta resolução dos modelos existentes na literatura, e é apontado um possível aumento artificial dos resultados quando não são consideradas múltiplas bases de dados para o treinamento. O modelo apresentado então é treinado com duas versões diferentes da base utilizada: no modelo proposto, usando classes de bases compostas, a acurácia de teste é de 91,21%; no modelo denominado como inflacionado, é utilizado um banco de dados diferente para cada classe, e a acurácia sobe para 98,33%, assim demonstrando os resultados aumentados artificialmente ao qual os autores criticam.

Em geral, o modelo apresentado também é bastante robusto, no entanto, assim como no trabalho de Wang et al. (2020), não são apresentadas informações sobre validação cruzada ou a média e desvio padrão das métricas utilizadas para a avaliação dele. Além disso, não é feita uma análise sobre as possíveis áreas de interesse detectadas pela RNA ou algum teste para verificar a capacidade de generalização do modelo em outras bases.

Quadro 1 – Comparativo entre publicações relevantes

Artigo	F1-Score	Acurácia	Metodologia
(CHAKRABORTY; MURALI; MITRA, 2022)	93,00%	96,43%	Segmentação de imagens Transfer learning Data augmentation Criação de nova arquitetura
(BHATTACHARYYA et al., 2022)	95,00%	96,60%	Data augmentation Segmentação de imagens Transfer learning Feature extraction convencional Criação de nova arquitetura

(TEIXEIRA et al., 2021)	88,00%	-	Segmentação de imagens Transfer learning
(NIKOLAOU et al., 2021)	90,00%	93,00%	Data augmentation Transfer learning
(MAIOR et al., 2021)	-	91,21%	Data Augmentation Criação de nova arquitetura
(NEFOUSSI; AMAMRA; AMAROUCHE, 2021)	94,00%	94,00%	Transfer learning
(UMER et al., 2021)	95,51%	89,85%	ImageDataGenerator Transfer learning
(RAJKUMAR et al., 2020)	-	96,00%	Segmentação de imagens Transfer learning Data augmentation
(KHAN; SHAH; BHAT, 2020)	95,60%	95,00%	Transfer Learning
(WANG; LIN; WONG, 2020)	-	93,30%	Transfer learning Criação de nova arquitetura
(ABBAS; ABDELSAMEA; GABER, 2020)	-	93,1%	DeTraC Transfer learning PCA Data augmentation
(UCAR; KORKMAZ, 2020)	98,25%	98,26%	Transfer learning Data augmentation Otimização bayesiana
(APOSTOLOPOULOS; MPESIANA, 2020)	-	94,72%	Transfer learning
(LOEY; SMARANDACHE; KHALIFA, 2020)	85,19%	85,19%	Transfer learning GAN

Fonte: própria

Teixeira et al. (2021), Rajkumar et al. (2020) e Chakraborty et al (2022) tentaram diminuir o viés dos conjuntos de dados utilizando a técnica de segmentação de imagens. O trabalho de RAJKUMAR et al. acionou profissionais da área da radiografia para realizar a segmentação manual das imagens, portanto, apresenta uma variedade de dados consideravelmente menor do que os outros dois trabalhos citados, os quais utilizam redes treinadas com imagens RX e suas respectivas máscaras de segmentação. O trabalho de Teixeira et al. (2021) é, em específico, interessante por trazer uma análise que compara a capacidade de generalização do modelo segmentado com o modelo não segmentado.

Em Bhattacharyya et al. (2022) também se utiliza a segmentação de imagens para diminuição do viés, no entanto, ele apresenta uma abordagem diferente dos citados anteriormente para o problema de classificação: nele há uma combinação e comparação de

técnicas de *deep learning* e aprendizado de máquina convencional. Uma análise comparativa do desempenho da classificação é realizada entre as diferentes arquiteturas propostas combinando redes profundas, métodos de extração de pontos-chave, e modelos de aprendizado de máquina. A maior precisão de classificação do modelo proposto pelo trabalho é de 96,6%, utilizando o modelo VGG-19 associado ao *Binary Robust Invariant Scalable Keypoints* (BRISK).

Apesar de realizar comparações bastante interessantes, a acurácia dos modelos apresentados pode estar aumentada, porque os autores, ao realizar a etapa do aumento de dados, não separam o conjunto de dados entre teste e treino antes das aplicações das técnicas de aumento de dados, causando, dessa forma, uma possível melhoria no desempenho do seu modelo.

Ucar e Korkmaz (2020) criticam o uso de arquiteturas pesadas e bases de dados privadas, sugerindo a utilização da rede *SqueezeNet* com a adição da otimização bayesiana, a fim de ajustá-la para o diagnóstico da COVID-19. Também é realizado aumento de dados, ou *data augmentation* (DA); porém, no *dataset*, é feito apenas para a classe COVID-19 antes da separação do treino e teste, dessa forma aumentando as métricas de avaliação do modelo, que apresenta um F1-score de 98,25% e acurácia de 98,26%.

Quadro 2 – Comparativo entre as bases de publicações relevantes

Artigo		**Base de Dados**	**Quantidade de fonte de dados diferentes**	**Quantidade de imagens de treino**
(CHAKRABORTY; MURALI; MITRA, 2022)		Cohen OCT CoronaHack-Chest	3	13251
(BHATTACHARYYA et al., 2022)		Cohen RSNA	2	3000
(TEIXEIRA et al., 2021)		Cohen RSNA ActualMed Figure 1 Radiopaedia Euroad Hamimi's Dataset Bontrager and Lampignano's Dataset	8	1727
(NIKOLAOU et al., 2021)		COVID Radiography Database	1	10606

(MAIOR et al., 2021)		COVIDx OCT	7	30544
(NEFOUSSI; AMAMRA; AMAROUCHE, 2021)		COVID-19 Radiography Database Cohen OCT	3	1089
(UMER et al., 2021)		Kaggle Covid-19 patients lungs x ray images 10000 Covid Data GradientCrescent	2	7000
(RAJKUMAR et al., 2020)		Cohen RSNA	2	12721
(KHAN; SHAH; BHAT, 2020)		Cohen OCT	2	1251
(WANG; LIN; WONG, 2020)		COVIDx	6	15599
(ABBAS; ABDELSAMEA; GABER, 2020)		JSRT Cohen	2	196
(UCAR; KORKMAZ, 2020)		COVIDx	6	3687
(APOSTOLOPOULOS; MPESIANA, 2020)		Cohen RSNA SIRM Radiopaedia OCT	5	1428
(LOEY; SMARANDACHE; KHALIFA, 2020)		Cohen OCT	2	8100

Fonte**:** Própria

Os trabalhos apresentados por Umer et al. (2021), Khan et al. (2020) e Abbas et al. (2020) com acurácias de 89,85%, 95,00% e 93.1%, respectivamente, utilizam *transfer learning* em conjuntos de dados com menos de 400 imagens para cada classe, entre 2 e 3 fontes de dados diferentes, assim, não contendo muita variedade de dados. Tanto Umer et al. (2021) quanto ABBAS et al. (2020) tentam contornar o problema da pouca quantidade de dados com técnicas de aumento de dados simples.

O trabalho de Nefoussi et al. (2021) que também utiliza a técnica de *transfer learning*, corresponde uma base de dados com mais de 6000 imagens, porém desbalanceada, visto que apenas 518 imagens pertencem a pacientes com COVID-19. Dessa forma, os autores optaram pela realização de um *undersampling* das classes para que apresentem a mesma quantidade de imagens cada, assim fazendo com que a CNN fosse treinada com aproximadamente 500 imagens por classe.

Loey, Smarandache e Khalifa (2020) tentaram contornar o problema da pequena quantidade de dados em 2020, no início da pandemia, utilizando GANs como alternativa ao aumento de dados convencional para gerar imagens novas baseadas num conjunto de dados limitado. Utilizando técnicas de *transfer learning* e da rede Alexnet, o modelo sugerido pelo artigo atinge 85,2% de acurácia de teste. Entretanto, neste trabalho, a etapa de geração de imagens é feita antes da separação de treino e teste, gerando viés no modelo.

O artigo de Apostolopoulos e Mpesiana (2020) também utiliza *transfer learning*, comparando o desempenho de 4 redes treinadas com 2 conjuntos de dados diferentes, o primeiro sendo uma coleção de 1428 imagens de RX, incluindo 224 imagens de COVID-19, 700 imagens com pneumonia bacteriana e 504 imagens de pulmões em condições normais. E, o segundo, um conjunto de dados incluindo 224 imagens de pacientes com COVID-19, 714 imagens com pneumonia bacteriana e viral e 504 imagens de pacientes em condições normais. Os dados foram coletados a partir de imagens de RX disponíveis em repositórios médicos públicos. Para o primeiro *dataset*, a melhor acurácia foi de 93,48% com a rede VGG-19; já para o segundo conjunto, a melhor acurácia atingida foi de 94,72% ao utilizar a rede *MobileNetv2*.

Nikolaou et al. (2021) desenvolveram um modelo via *transfer learning*, utilizando a *EfficientNetB0* com um banco de dados para treinamento com 15.153 imagens de RX. Também usaram o aumento de dados para evitar o fenômeno do *overfitting*, com o objetivo de solucionar o desequilíbrio de classes. Foi criado um modelo robusto cuja acurácia de teste é de 93,00%. Entretanto, dados experimentais sobre as médias e desvios padrões das métricas utilizadas não são disponibilizados, e, apesar do trabalho utilizar uma base de dados externa para avaliar a capacidade de generalização do modelo proposto, não é explicado de onde os dados externos foram coletados, nem quantas imagens externas foram utilizadas para avaliação do modelo.

Ao realizar a análise dos artigos, nota-se poucos estudos que avaliam a capacidade de generalização dos modelos criados e que metodologicamente existem falhas ou falta de informação em relação ao número de repetições dos experimentos. Em sua grande maioria, os trabalhos apresentam o melhor resultado que obtiveram, e não resultados médios e seus desvios padrões, possivelmente apresentando resultados aumentados e modelos performados de maneira inferior na vida real. Também há poucos estudos que comparam diversas técnicas ou treinamentos com bases de dados diferentes ao avaliar se há melhoria no desempenho de teste e na capacidade de generalização.

Portanto, o presente trabalho tem como objetivo avaliar, utilizando a metodologia científica, o desempenho e a capacidade de generalização das combinações de 3 técnicas já utilizadas pela literatura: a segmentação de imagens, o *transfer learning* e o aumento de dados

por meio de redes generativas adversariais, em 2 conjuntos de dados coletados provenientes de 8 bases de dados abertas, totalizando 5130 imagens para treino no primeiro conjunto de dados, e 6471 imagens de treino no segundo conjunto de dados utilizado. Este estudo procura, então, responder as seguintes perguntas: é possível encontrar uma *pipeline* ótima para classificação das doenças respiratórias (Covid-19, Pneumonia ou ausência de doença)? qual *pipeline* performa melhor em quesito de generalização?

2 FUNDAMENTAÇÃO TEÓRICA E REVISÃO DA LITERATURA

Esse capítulo almeja embasar conceitos referentes ao aprendizado de máquina, aprendizado profundo e redes neurais, como também explanar as arquiteturas das redes utilizadas nesta pesquisa. Em seguida será apresentada também uma breve visão geral sobre a utilização de aprendizado de máquina em problemas de classificação de imagens médicas.

2.1 APRENDIZADO PROFUNDO

O aprendizado profundo (ou *deep learning*) refere-se a uma família de algoritmos de aprendizado de máquina que fazem uso intensivo de redes neurais artificiais. Como resultado, é também conhecido como redes neurais profunda, ou *Deep Neural Networks* (DNNs). Em essência, o aprendizado profundo transforma e extrai características através de uma cascata de muitas camadas de unidades de processamento não-lineares. As camadas superiores aprendem características mais complexas derivadas de características de camadas inferiores, enquanto as camadas inferiores adjacentes à entrada de dados aprendem características mais simples. Uma representação hierárquica e potente das características é criada por essa arquitetura.

Modelos computacionais com numerosas camadas de processamento podem aprender representações de dados em vários níveis de abstração devido ao aprendizado profundo. O estado da arte em muitos outros campos, incluindo descoberta de drogas e genética, identificação de objetos, reconhecimento visual de objetos e reconhecimento de fala, foi significativamente aprimorado por estas técnicas. O aprendizado profundo identifica estruturas complexas em grandes conjuntos de dados, utilizando a técnica de retropropagação para sugerir como uma máquina deve modificar seus parâmetros internos que são usados para calcular a representação em cada camada a partir da representação na camada anterior. As redes recorrentes têm esclarecido problemas com tipos de dados sequenciais como texto e fala, enquanto as redes convolucionais profundas fizeram avanços no processamento de imagens, vídeo, fala e áudio (LECUN; BENGIO; HINTON, 2015).

Embora os conceitos subjacentes às DNNs já existam há algum tempo, não foi até os anos 2010 que eles realmente começaram a decolar em termos de uso. Com base em descobertas revolucionárias em Dahl et al. (2011), o campo do reconhecimento automático da fala foi o primeiro a aplicar estas técnicas. Esse método foi rapidamente adotado tanto pela academia quanto pelas empresas, tornando-se o paradigma aceito (HINTON et al., 2012). A conquista que mais chamou a atenção, entretanto, foi quando Krizhevsky, Sutskever e Hinton (2012)

demonstraram que CNNs profundas poderiam melhorar significativamente o desempenho no difícil *benchmark* de classificação de imagens *ImageNet* (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), reduzindo a taxa de erro de 26% para 16% em um único ano. Isso foi uma melhoria significativa em relação à taxa anterior de progresso de cerca de 2% de redução por ano.

2.2 APRENDIZADO PROFUNDO APLICADO ÀS IMAGENS MÉDICAS

O interesse pelo aprendizado profundo pode ser visto nas inúmeras chamadas, desafios, conferências ou nos resultados que todos os grupos de pesquisa ao redor do mundo apresentam. Devido a sua adaptabilidade, alto desempenho, grande capacidade de generalização e diversos usos, entre muitos outros atributos, o mundo científico tem concentrado sua atenção na aprendizagem profunda. As numerosas contribuições tornam os modelos atuais cada vez mais eficazes, e novos avanços são publicados regularmente (WASON, 2018). A produção de dados digitais em larga escala, fortes infraestruturas computacionais, unidades de processamento gráfico, ou *graphic processing units* (GPU), e computação em nuvem contribuíram para a expansão do aprendizado profundo em vários domínios científicos, incluindo a medicina.

Para a situação específica das imagens médicas, o aprendizado profundo é normalmente aplicado para a resolução de uma diversa gama de problemas, incluindo os citados a seguir: classificação de imagens, como determinar se um paciente apresenta uma doença ou a gravidade de determinada doença (RAJPURKAR et al., 2017b); detecção de objetos, em que uma de suas aplicações identificam massas ou tumores a partir de órgãos (LI et al., 2019); segmentação de imagens, que é usada principalmente para segmentar pulmões, tumores, discos ópticos, nódulos e imagens cardíacas (WANG et al., 2020); e geração de imagens, que pode gerar dados sintéticos para síntese de ressonância magnética ou mais dados de TC ou RX quando há necessidade disso, ou seja, quando o conjunto de dados de treinamento é muito pequeno (SHORTEN; KHOSHGOFTAAR, 2019).

Mesmo com muitos resultados encorajadores de pesquisas anteriores, ainda há uma série de problemas a serem resolvidos antes que o aprendizado profundo seja devidamente aplicado às imagens médicas. Por exemplo, há uma alta dependência da quantidade e qualidade dos conjuntos de dados de treinamento, bem como uma tendência para o *overfitting* e o enviesamento. Isso significa que a disponibilidade de dados distorce o desempenho e restringe as implementações a um subconjunto com traços particulares (ANAYA-ISAZA; MERA-JIMéNEZ; ZEQUERA-DIAZ, 2021).

Uma generalização do aprendizado profundo deve ser feita dadas as variações na prevalência de doenças, na modalidade de imagens e nos protocolos do ambiente clínico em todo o mundo. Com o objetivo de verificar a eficácia de cada metodologia, os métodos de avaliação também devem ser melhorados, considerando o aspecto de caixa preta dos algoritmos de aprendizagem profunda existentes. Mesmo quando uma abordagem baseada no aprendizado profundo produz resultados excepcionais, pode ser frequentemente desafiador ou impossível articular o raciocínio por trás de uma escolha.

Entretanto, os médicos agora têm que lidar com leituras cada vez mais complexas. Torna-se um desafio completar a leitura de uma imagem dentro do prazo e produzir relatórios relevantes como resultado. Ao fornecer uma análise quantitativa das lesões preocupantes ou classificação prévia de uma doença, por exemplo, espera-se que o aprendizado profundo ajude os radiologistas a fazer um diagnóstico mais preciso. Ele também pode permitir um processo clínico mais rápido (KIM et al., 2019).

A técnica de *deep learning* já demonstrou um desempenho em tarefas de reconhecimento e visão computacional comparável ao dos humanos (DODGE; KARAM, 2017). É pertinente assumir que as práticas de saúde podem sofrer algumas modificações significativas como resultado desses avanços tecnológicos. Em vez de substituir os médicos, quando se trata da aplicação da IA em imagens médicas, é possível afirmar que esse avanço tecnológico funcionará como uma ferramenta colaborativa para reduzir a carga e a distração de muitas tarefas rotineiras e repetitivas.

Um conhecimento detalhado da tecnologia da IA entre médicos e cientistas ou engenheiros da computação, assim como a prática clínica e o fluxo de trabalho mais adequados, seria um dos elementos cruciais para o desenvolvimento e uso clínico adequado da IA para a medicina. O desenvolvimento da IA para a utilização de dados de quadros clínicos também deve levar em consideração uma série de outras dificuldades, tais como aquelas que devem ser resolvidas e conquistadas nos âmbitos ético, legais e regulatórios.

2.3 INTERPRETABILIDADE DE MODELOS DE APRENDIZADO PROFUNDO

Em geral, com o aumento da sua complexidade e precisão, os modelos de aprendizagem de máquina tendem a tornar-se complexos e, portanto, difíceis de interpretar (KATUWAL; CHEN, 2016). Modelos com estruturas de fácil compreensão e um número limitado de parâmetros, tais como regressão linear ou árvores de decisão, podem, geralmente, ser interpretados sem necessidade de algoritmos de explicação adicionais. Em contraste, modelos

complexos, tais como modelos de aprendizado profundo, são considerados caixas pretas, porque seu comportamento não pode ser compreendido, mesmo conhecendo sua estrutura e pesos.

Para aplicações em problemas reais, principalmente no âmbito da saúde, é importante compreender o processo de decisão de um modelo preditivo antes que a ela possa ser útil. Portanto, um modelo preditivo tem de ser interpretável, ou transformado para ser interpretável, para que o utilizador do modelo possa compreender o seu processo de decisão (KATUWAL; CHEN, 2016). Uma das formas de adicionar interpretabilidade a tais modelos é por meio de *local surrogate models*. O LIME é um modelo comumente utilizado por pesquisadores para adicionar interpretabilidade a modelos caixa preta (DIEBER; KIRRANE, 2020).

2.3.1 LIME

O LIME pode ser classificado como um *local surrogate model*, ou modelo local de substituição. É um modelo interpretável usado para explicar previsões individuais de modelos de aprendizagem de máquina caixa preta. O LIME é um trabalho no qual os autores propõem uma implementação concreta de modelos locais de substituição. Os modelos substitutos são treinados para aproximar as predições do modelo de caixa preta subjacente. E, ao invés de treinar um modelo substituto global, o LIME se concentra no treinamento de modelos substitutos locais para explicar as previsões individuais.

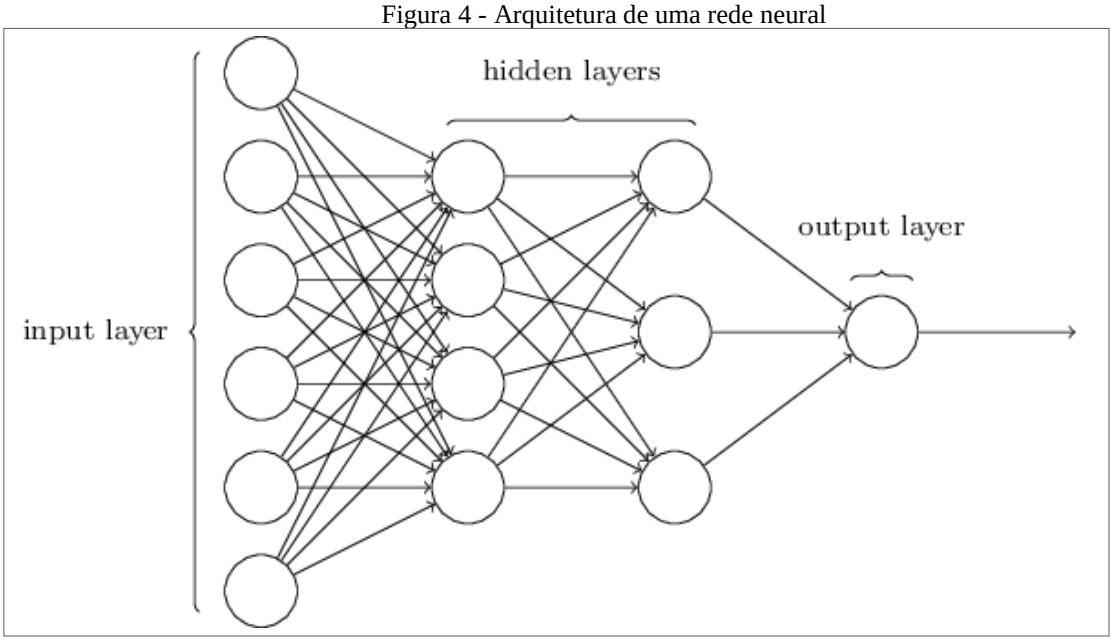
A ideia geral por trás desses modelos é bastante intuitiva: o objetivo é entender a razão pela qual o modelo de aprendizagem de máquina associado fez certa previsão. O LIME testa o que acontece com as previsões quando variações de seus dados no modelo original são criados. Assim, é gerado um novo conjunto de dados que consiste em amostras distorcidas e previsões correspondentes ao modelo caixa preta. Neste novo conjunto de dados, o LIME treina um modelo interpretável, que é ponderado pela proximidade das instâncias amostradas com a instância de interesse (RIBEIRO; SINGH; GUESTRIN, 2016).

2.4 REDES NEURAIS

As redes neurais, também conhecidas como redes neurais artificiais, ou *Artificial Neural Networks* (ANNs), e redes neurais simuladas, ou *Simulated Neural Networks* (SNNs), são um subconjunto de aprendizagem de máquina e estão no centro dos algoritmos de aprendizagem profunda. Sua estrutura e nomenclatura são modeladas conforme o cérebro humano, utilizando

neurônios artificiais para espelhar a comunicação entre os neurônios orgânicos. Elas são compostas de camadas de nós, cada qual com uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (NIELSEN, 2015). Tal arquitetura pode ser visualizada na figura 4.

Figura 4 - Arquitetura de uma rede neural



Fonte: (NIELSEN, 2015)

A camada de entrada recebe as características de entrada. Nenhum cálculo é feito nessa camada; os nós aqui apenas enviam informações (características) para uma camada oculta. Ele dá à rede informações do mundo exterior.

Na camada oculta, os nós não são expostos ao mundo exterior, eles são a parte da abstração proporcionada por qualquer rede neural. A camada oculta faz vários cálculos aos recursos inseridos através da camada de entrada e envia os resultados para a camada de saída, que comunica o conhecimento adquirido pela rede para o mundo exterior. A Figura 4 mostra um *layout* de rede neural simples de quatro camadas.

Dentro de uma camada, cada nó, ou neurônio artificial, está conectado a outro e tem um peso, um limiar e uma função de ativação correspondentes. A função de ativação calcula um total ponderado e depois adiciona um viés a ele para determinar se um neurônio deve ou não ser ativado. O objetivo principal da função de ativação é adicionar não-linearidade à saída de um neurônio (HAYKIN, 2008). Um nó é ativado e começa a transferir dados para a próxima camada da rede se sua saída for maior do que o valor limite pré-definido para aquele nó. Caso contrário, nenhum dado é transmitido para a camada seguinte da rede.

É sabido que os neurônios em redes neurais se comportam de acordo com o peso, o viés e suas respectivas funções de ativação. A retropropagação é o processo de atualização dos pesos e dos vieses dos neurônios em uma rede neural com base no erro na saída (MURPHY, 2022). A retropropagação é possibilitada pelas funções de ativação, pois elas fornecem os gradientes e erros necessários para atualizar os pesos e os enviesamentos.

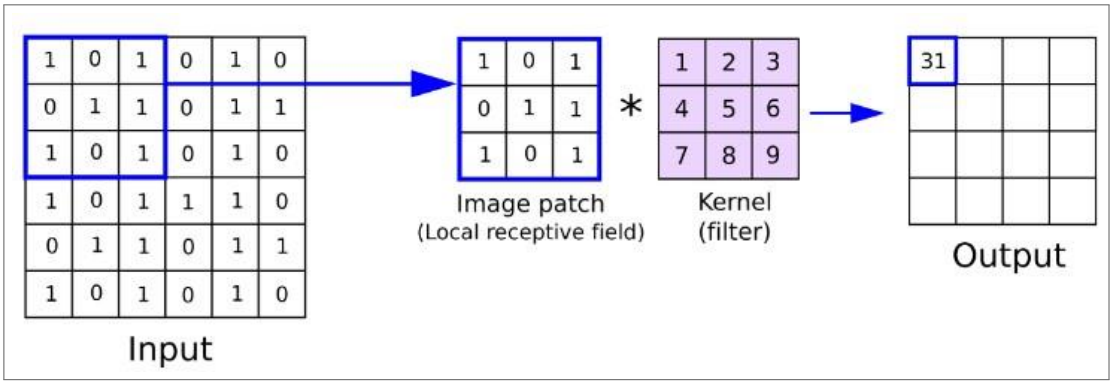
Os algoritmos de aprendizagem de máquinas baseados em redes neurais normalmente não precisam ser projetados com regras particulares que especifiquem o que se deve prever a partir da entrada. Em vez disso, o algoritmo de aprendizagem da rede neural aprende analisando uma grande quantidade de amostras rotuladas (ou seja, dados que foram fornecidos com "respostas") e utilizando esta chave de resposta para determinar quais qualidades da entrada são necessárias para criar a saída desejada. Após processar um número suficiente de exemplos, a rede neural pode começar a processar insumos novos, não testados e produzir resultados confiáveis. Como o computador aprende por experiência, quanto mais exemplos e entradas diversas ele vê, mais precisos são os resultados geralmente obtidos.

2.4.1 Redes Neurais Convolucionais

Dentro do campo das redes neurais existem as redes convolucionais, comumente chamadas de redes neurais convolucionais (CNNs). Essas são uma classe particular de rede neural utilizadas para processar dados com uma arquitetura tipo grade específica (LECUN et al., 1989). Exemplos de tal arquitetura incluem dados de imagem, podendo ser mostrados como uma grade 2-D de *pixels*, e dados de série temporal, que conseguem ser visualizados como uma grade 1-D coletando amostras em intervalos regulares.

Redes convolucionais alcançaram grande sucesso em diversos cenários do mundo real. O termo "rede neural convolucional" refere-se a uma rede que utiliza a operação matemática da convolução, ilustrada na figura 5. Elas são essencialmente redes neurais com pelo menos uma camada que utiliza a convolução em vez da multiplicação matricial padrão (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 5 – Operação de convolução



Fonte: (REYNOLDS, 2017)

A camada de convolução é a primeira a extrair características de uma imagem de entrada. Uma convolução preserva a relação entre os *pixels* através da aprendizagem das características da imagem usando pequenos quadrados de dados de entrada. É uma operação matemática que leva duas entradas, podendo ser a matriz de imagem e um filtro ou *kernel*. Um filtro se refere a uma pequena matriz, e o operador de convolução dá origem a uma nova imagem em que cada elemento é uma combinação ponderada das entradas de uma região ou *patch* da imagem original (também chamado de campo receptivo local) com pesos dados pelo filtro. Desse modo, uma convolução é um produto de pontos de duas matrizes achatadas: um *kernel* e um *patch* de uma imagem do mesmo tamanho. A convolução de uma imagem com diferentes filtros pode realizar operações como detecção de borda, desfocagem e nitidez.

Dentre as redes neurais convolucionais existentes, três serão detalhadas neste capítulo por serem as redes escolhidas para a etapa experimental do presente trabalho. Sendo elas a *VGG-16*, *Inception-ResNet-v2* e *DenseNet-121*.

2.4.1.1 *VGG-16*

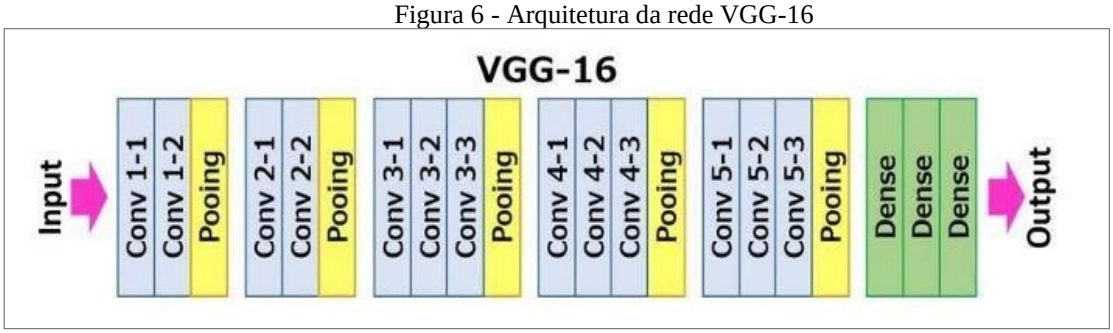
Em seu artigo “Very Deep Convolutional Networks for Large-Scale Image Recognition”, Simonyan e Zisserman (2014), da Universidade de Oxford, apresentaram o modelo de rede neural convolucional conhecido como VGG-16. A precisão do teste top-5 para o modelo na ImageNet, um conjunto de dados com mais de 14 milhões de imagens divididas em 1000 classes, é de 92,7%. Ele se tornou um modelo bastante conhecido ao ser submetido ao ImageNet Large Scale Visual Recognition Challenge 2014 (ILSVRC-2014), configurando-se como um dos principais modelos de visão computacional da atualidade.

Os desenvolvedores desse modelo analisaram as redes concebidas anteriormente e aumentaram a profundidade usando uma arquitetura com filtros de convolução bastante

pequenos (3 por 3), o que demonstrou um avanço notável sobre as configurações anteriores de estado da arte.

Na VGG-16, a primeira camada convolucional recebe uma imagem RGB de tamanho fixo com dimensões de 224 por 224. Após isso, a imagem é colocada através de uma série de camadas convolucionais cujos filtros utilizam um campo receptivo estreito (3 por 3), sendo o menor tamanho para capturar os conceitos de esquerda, direita, superior, inferior e centro. Em uma das configurações, são usados filtros de convolução 1 por 1, que podem ser pensados como uma transformação linear dos canais de entrada (seguidos pela não linearidade). O salto (*stride*) de convolução é fixado em 1 pixel; o acolchoamento (*padding*) espacial da entrada de camadas convolucionais representa a resolução espacial preservada após a convolução, ou seja, o *padding* é de um pixel para camadas convolucionais 3 por 3. O agrupamento espacial (*spatial pooling*) é realizado por cinco camadas de *max-pooling*, que acompanham algumas das camadas convolucionais (nem todas as camadas são seguidas por ele). Com salto de dois, o *max-pooling* é realizado sobre uma janela de 2 por 2 pixels. A arquitetura descrita pode ser visualizada na figura 6.

Figura 6 - Arquitetura da rede VGG-16



Fonte: (LEARNING, 2021)

Seguindo uma série de camadas convolucionais (que variam em profundidade entre as arquiteturas), três delas totalmente conectadas são utilizadas: as duas primeiras têm 4.096 canais cada, enquanto a terceira realiza a classificação ILSVRC e, assim, tem 1.000 canais (um para cada classe). A camada *soft-max* é a última, e a unidade linear retificada (ReLu, abreviação para *rectified linear unit*) é a função de ativação presente em todas as camadas ocultas da VGG-16.

2.4.1.2 *DenseNet-121*

Cada camada convolucional de uma rede neural convolucional convencional, além de receber a entrada da primeira, recebe a saída da camada anterior e gera um mapa de

características de saída, que é então passado para a camada convolucional seguinte. Como resultado, existem conexões “L” diretas para cada “L” camadas, uma entre cada uma delas e a seguinte.

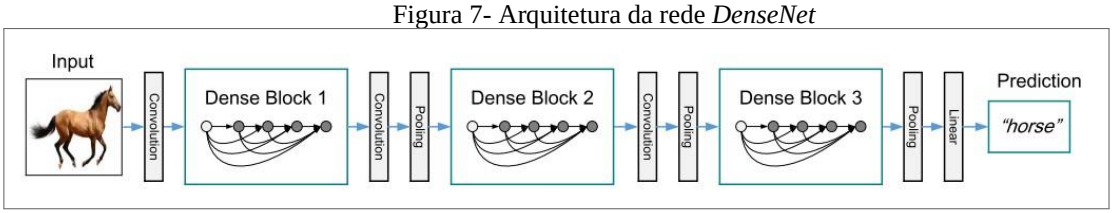
O problema do “gradiente de fuga” se desenvolve quando a CNN começa a ter mais camadas, quando se torna mais profunda. Isto é, no momento em que a transmissão de informações das camadas de entrada para as de saída se prolonga, algumas informações podem “desaparecer” ou “se perder”, o que diminui a capacidade da rede de aprender de forma eficaz.

Ao alterar a arquitetura típica da CNN e otimizar a conectividade entre as camadas, a *DenseNet* ameniza esse problema. Cada camada está ligada a todas as outras camadas em uma arquitetura *DenseNet*, dando origem ao termo “rede convolucional densamente conectada”. Existem L(L+1)/2 conexões diretas para L camadas. Todos os mapas de características das camadas anteriores são utilizados como entradas para cada uma delas, e os próprios mapas de características dessa camada são utilizados como entradas para cada uma adicional.

Os mapas de características de todas as camadas anteriores não são somados, mas sim concatenados e usados como insumos em cada uma delas. Como resultado, *DenseNets* precisam de menos parâmetros do que uma CNN padrão comparável, o que permite a reutilização de recursos porque os mapas de características duplicados são eliminados (HUANG et al., 2016).

Além das camadas fundamentais de convolução e *pooling*, *DenseNets* também incluem dois elementos cruciais de construção. Eles são as camadas de transição e os blocos densos. Os blocos densos são estruturas em que o tamanho dos mapas de características dentro de um bloco é fixo, mas o número de filtros entre eles é variável. As camadas de transição ocorrem entre os blocos, elas cortam o número de canais pela metade.

Figura 7- Arquitetura da rede *DenseNet*



Fonte: (HUANG et al., 2016)

Uma DenseNet profunda com três blocos densos é exibida na figura 7. Enquanto o tamanho dos mapas de características dentro do bloco denso é o mesmo para permitir a concatenação de características, as camadas de transição entre dois blocos adjacentes executam *downsampling* (ou seja, mudam o tamanho dos mapas de características) através de processos de convolução e *pooling*.

Figura 8 - Arquitetura completa da rede *DenseNet*

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Fonte**:** (HUANG et al., 2016)

Ao analisar a figura 8, a qual representa a arquitetura completa da *DenseNet-121*, observa-se que cada bloco denso possui um número diferente de camadas (repetições), em que cada uma das quais contém duas convoluções: uma camada de gargalo que é 1 por 1 em tamanho, e uma camada de convolução que é 3 por 3 em tamanho.

Uma camada convolucional de 1 por 1 e outra média de 2 por 2 com um passo de dois também estão presentes em cada camada de transição. Sendo assim as que estão presentes na *DenseNet-121* são as seguintes:

- 1 camada de convolução com 64 filtros de tamanho 7 por 7 e pulo de tamanho 2;
- 58 camadas de convolução com 3 por 3 de tamanho;
- 61 camadas 1 por 1 de convolução;
- 4 camadas de *average pooling*;
- 1 camada totalmente conectada.

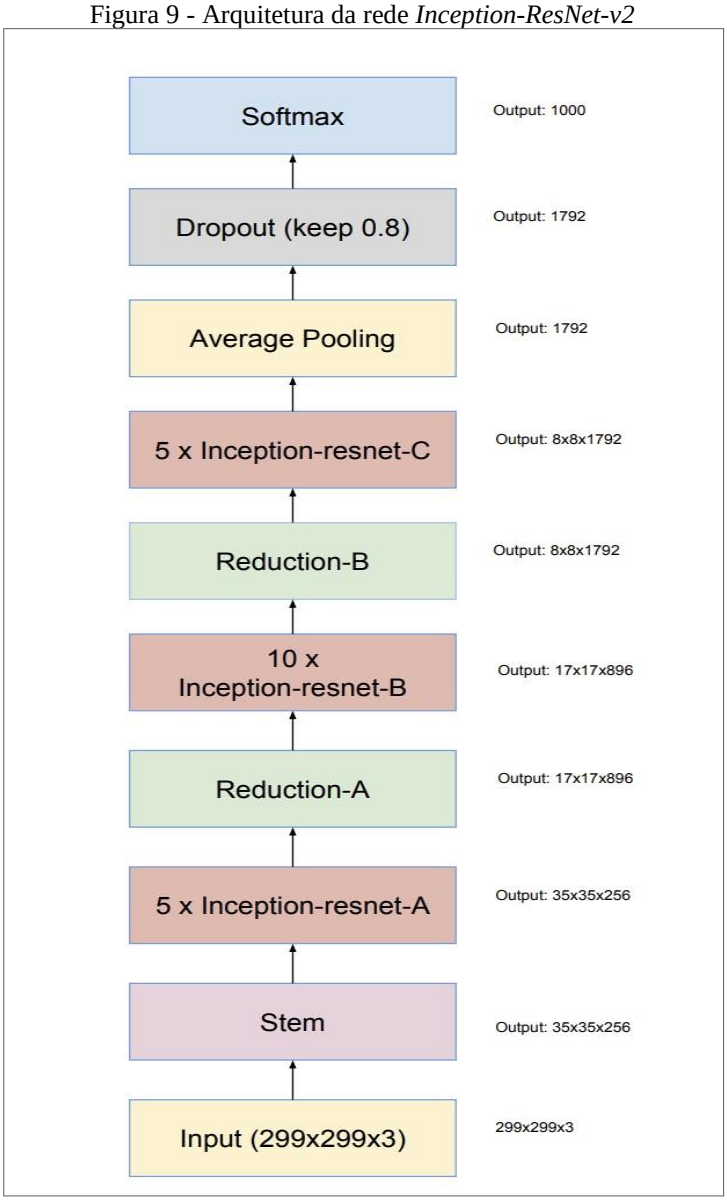
Dessa forma, a *DenseNet-121* possui 120 convoluções e quatro *average poolings*.

2.4.1.3 *Inception-ResNet-v2*

Com um desempenho muito forte a um custo computacional baixo, as arquiteturas *ResNet* e *Inception* têm estado no centro das maiores melhorias no desempenho de reconhecimento de imagem nos últimos anos. A *Inception-ResNet-v2* é uma rede neural convolucional que combina os conceitos de ambas as arquiteturas anteriores, sendo formulada acima da estrutura da *Inception* combinada com a conexão residual. No bloco *Inception-ResNet*,

filtros convolucionais de múltiplos tamanhos são combinados com conexões residuais. O uso de conexões residuais não só evita o problema de degradação causado por estruturas profundas, mas também reduz o tempo de treinamento. A Figura 9 mostra a arquitetura básica de rede da *Inception-ResNet-v2*.

Figura 9 - Arquitetura da rede *Inception-ResNet-v2*



Fonte: (SZEGEDY et al., 2016)

Indo mais a fundo, a ideia por trás da *Inception-ResNet-v2* era de aumentar a eficiência computacional e diminuir o gargalo de representação. Para esse segundo objetivo, a suposição era que as redes neurais operam mais eficazmente quando as convoluções não alteram significativamente as dimensões da entrada. Um "gargalo representacional" ou

representational bottleneck seria uma perda de informação que pode ocorrer quando as dimensões são reduzidas em demasia.

A solução desses problemas, proposta por Szegedy et al. (2016), foi de utilizar métodos inteligentes de fatorização, combinação de convoluções e expansão de filtros. Assim, as convoluções tornam-se mais eficientes em termos de complexidade computacional e diminuindo o gargalo representacional. Na medida em que se fatoriza as convoluções 5 por 5 para duas operações 3 por 3, se obtém uma melhora na velocidade computacional. Mesmo que pareça contraintuitivo, uma convolução 5 por 5 é 2,78 vezes mais cara do que uma convolução 3 por 3, portanto, o empilhamento de duas convoluções 3 por 3 leva a um aumento no desempenho.

Além disso, Szegedy et al. (2016) combinam convoluções 1xn e nx1 com convoluções nxn de tamanho de filtro. Uma convolução 3 por 3, por exemplo, é equivalente a primeiro executar uma 1 por 3 e depois aplicar outra 3 por 1 à saída. Eles descobriram que sua abordagem era 33% mais barata do que uma única convolução 3 por 3.

A fim de eliminar o gargalo representacional, os bancos de filtros nos módulos da rede foram aumentados (ampliados ao invés de aprofundados). O módulo perderia informações se fosse aumentada a profundidade dos filtros, pois as dimensões seriam drasticamente reduzidas. Três tipos diferentes de módulos iniciais (chamados de módulos A, B e C a partir da ordem de criação) foram construídos usando os conceitos acima mencionados. Eles também podem ser vistos na arquitetura representada na figura 9.

2.5 AUMENTO DE DADOS

Há numerosos campos acadêmicos que utilizam redes profundas e convolucionais para lidar com problemas de visão computacional em um esforço para superar os *benchmarks* mais atuais. Um dos problemas mais difíceis é melhorar a capacidade de generalização desses modelos. A generalizabilidade mede a diferença entre o desempenho de um modelo em dados já vistos anteriormente (dados de treinamento) com seu desempenho em dados que nunca viu antes (dados de teste). Modelos pouco generalizáveis geralmente têm um problema de *overfit* ao conjunto de treinamento.

Uma das várias estratégias que podem ser usadas para diminuir o sobreajuste é o aumento de dados (PEREZ; WANG, 2017). No mundo real, pode-se ter um conjunto de dados de imagens tiradas de um conjunto específico de circunstâncias. Mas existem muitas situações distintas nas quais as imagens podem estar presentes, incluindo variações na escala, brilho,

orientação e posição. Ao fornecer a nossa rede neural dados referentes a apenas uma, ou algumas das situações possíveis, é factível que a rede não consiga realizar uma boa generalização para circunstâncias diferentes. Ao gerar dados adicionais alterados artificialmente durante o treinamento, pode-se levar em consideração essas circunstâncias. Os dados sintéticos representarão uma gama mais ampla de conjuntos de dados potenciais.

Dessa maneira, o aumento de dados se aproxima da raiz do problema: o conjunto de dados de treinamento. Isso é feito com a expectativa de que os aumentos permitirão a extração de mais informações do conjunto de dados original. Essas ampliações aumentam artificialmente o tamanho do conjunto de dados de treinamento através de distorções ou sobreamostragem de dados (SHORTEN; KHOSHGOFTAAR, 2019). As ampliações de distorção de dados transformam as imagens existentes de forma que seu rótulo seja preservado, isso abrange ampliações tais como transformações geométricas e de cor, apagamento aleatório, adição de ruído e embaçamento da imagem.

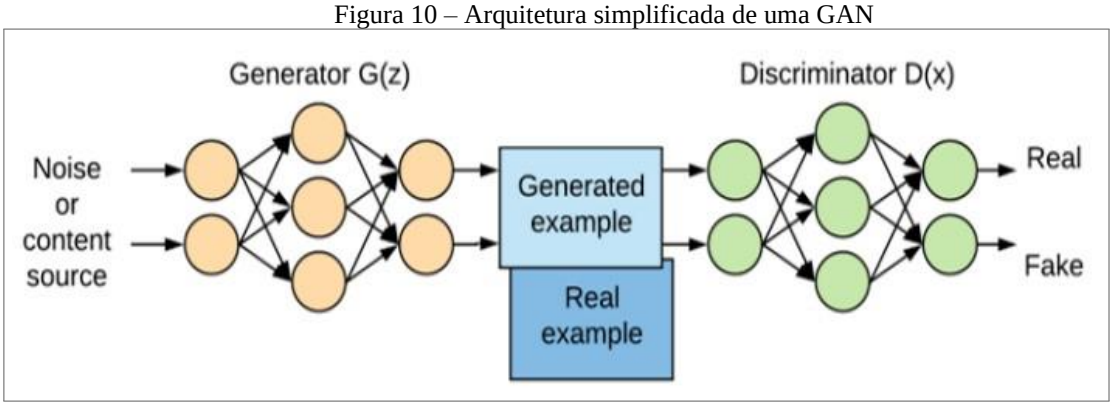
As ampliações de sobreamostragem criam instâncias sintéticas e as adicionam ao conjunto de treinamento, o que inclui a mistura de imagens e geração de novas delas por redes de adversárias generativas (GANs). Os aumentos de sobreamostragem e de distorção de dados não formam uma dicotomia mutuamente exclusiva. Por exemplo, amostras criadas por uma GAN podem ser empilhadas com cortes ou *zooms* aleatórios para aumentar ainda mais o conjunto de dados.

2.5.1 Redes generativas adversárias

Dentro dos métodos de aumento de dados, as redes generativas adversariais são uma tecnologia relativamente nova e potente que fazem a geração não supervisionada de novas imagens usando o método min-max. Desde sua introdução em 2014 por Goodfellow et al. (2014), as GANs foram descobertas como incrivelmente úteis em uma variedade de problemas de geração e manipulação de imagens, incluindo super-resolução (criar uma imagem de alta resolução a partir de uma de baixa resolução), restauração de pedaços ausentes de uma imagem, tradução imagem-imagem (por exemplo, conversão de esboços para imagens), síntese texto-imagem e mistura de imagens (combinando partes selecionadas de duas imagens para criar uma nova) (MIKOLAJCZYK; GROCHOWSKI, 2018). O excelente desempenho das GANs tem atraído mais atenção para a questão de como utilizá-las na tarefa de aumento de dados. Essas redes têm a capacidade de produzir novos dados sintéticos de treinamento, o que pode levar a modelos de classificação que têm um melhor desempenho.

Em geral, as GANs utilizam duas redes adversárias (G(z) e D(x)), em que uma delas cria uma imagem foto realista em um esforço para enganar a outra rede (gerador G(z)), sendo ela ensinada a diferenciar as imagens falsas das reais (discriminador D(z)). Dito de outra forma, o propósito do gerador é minimizar uma função de custo (valor) V(D,G) (por exemplo, probabilidade máxima), enquanto a tarefa do discriminador é maximizá-la (GOODFELLOW et al., 2014). A Figura 10 serve como um exemplo desta arquitetura.

Figura 10 – Arquitetura simplificada de uma GAN



Fonte: (MIKOLAJCZYK; GROCHOWSKI, 2018)

As redes geradoras e discriminadoras do modelo padrão da GAN utilizam redes *perceptron* multicamadas. Isso pode gerar imagens aceitáveis a partir de um conjunto de dados de imagem simples, como, por exemplo, os dígitos manuscritos do MNIST (DENG, 2012). Para conjuntos de dados mais complexos e de maior resolução, ela não produz resultados de alta qualidade (MIKOLAJCZYK; GROCHOWSKI, 2018).

Diversos estudos que alteram a estrutura da GAN utilizando diferentes designs de rede, funções de perda, técnicas evolutivas e outras técnicas têm sido publicados. A qualidade das amostras produzidas pelas GANs tem aumentado muito como resultado destas pesquisas. Muitas novas arquiteturas foram propostas para desenvolver a ideia das GANs e gerar imagens de saída com melhor resolução; entre estas novas arquiteturas, DCGANs, GANs de crescimento progressivo, CycleGANs e GANs Condicionais parecem ter o maior potencial para uso no aumento de dados. Como a DCGAN é a arquitetura utilizada nos experimentos deste trabalho ela será mais bem detalhada a seguir.

2.5.1.1 *DCGAN*

Em contraste com a arquitetura inicial das GANs, que usava apenas 5 camadas completamente conectadas, as DCGANs ou GANs convolucionais profundas são as primeira a

usar camadas convolutivas (RADFORD; METZ; CHINTALA, 2015). Devido a sua facilidade de uso, a DCGAN é frequentemente utilizada como a GAN *baseline* de fato. Ao mesmo tempo em que oferecem algumas sugestões úteis sobre o projeto da rede (uso de convoluções com saltos em vez de camadas de *pooling*, uso substancial da normalização em *batch*, etc.), as DCGANs demonstraram uma melhoria significativa na qualidade da imagem e estabilidade do treinamento.

A complexidade intrínseca das redes geradoras e discriminadoras foi expandida com a proposta da arquitetura da DCGAN, sendo essa estratégia essencial para produzir imagens com alta resolução. Em vez de utilizar *perceptrons* multicamadas, as CNNs são usadas como redes geradoras e discriminatórias nessa arquitetura. A DCGAN criada por Radford; Metz; Chintala (2015) foi treinada em vários conjuntos de imagens, entre eles a base de dados das imagens de interiores de quartos, LSUN (YU et al., 2015), com cada imagem medindo 64 por 64 por 3, sendo assim, significativamente maiores do que as imagens utilizadas no *dataset* MNIST. Os outros dois conjuntos de dados utilizados foram *Imagenet-1k* (DENG et al., 2009) e uma nova base de dados contendo rostos humanos, recentemente montada pelos autores.

Dessa forma, Radford; Metz; Chintala (2015) propõem um conjunto mais estável de arquiteturas para o treinamento de redes generativas adversariais e apresentam provas de que as redes com sua nova arquitetura aprendem boas representações de imagens para o aprendizado supervisionado e modelagem generativa. Ainda assim, existem algumas instabilidades no modelo proposto, principalmente quando são treinados por mais tempo.

2.6 SEGMENTAÇÃO DE IMAGENS

Outra técnica utilizada para melhorar a generalização de modelos em problemas de visão computacional é a segmentação de imagens, sendo utilizada para dividir uma imagem digital em várias partes de acordo com as diversas características dos *pixels*. Como a informação espacial de uma imagem é crucial para segmentar semanticamente várias regiões, muitas vezes é uma tarefa de visão de baixo nível ou similar ao *pixel*, em oposição à classificação e ao reconhecimento de objetos.

Com objetivo de facilitar a análise, a segmentação procura extrair informações úteis. Nesse caso, os *pixels* da imagem são rotulados para que cada *pixel* compartilhe qualidades específicas, tais como cor, intensidade e textura. Segmentação semântica (um problema de classificação de *pixels* com rótulos semânticos) e segmentação de instância (partição de objetos individuais) são os dois principais tipos de segmentação de imagem. Além disso, existe um

terceiro tipo de segmentação conhecida como segmentação panóptica, que combina os métodos dos dois primeiros (SULTANA; SUFIAN; DUTTA, 2020).

Para cada pixel na imagem, a segmentação semântica faz rotulagem no âmbito do pixel e se utiliza de um conjunto de categorias de objetos (por exemplo, humano, automóvel, árvore, céu), tornando-a uma tarefa mais difícil do que a classificação da imagem, que prevê um único rótulo para a imagem completa. Ao identificar e separar cada objeto de interesse na imagem (como a divisão de pessoas distintas), a segmentação de instância amplia o escopo da segmentação semântica (MINAEE et al., 2020).

Nos últimos anos, uma variedade de técnicas de segmentação tem sido desenvolvida, desde a segmentação de imagens em MATLAB até técnicas convencionais de visão computacional e técnicas de aprendizado profundo de última geração. A segmentação de imagens avançou significativamente, particularmente com o desenvolvimento de Redes Neurais Profundas.

Uma grande variedade de aplicações práticas de visão computacional, tais como a identificação de sinais de trânsito, biologia, avaliação de materiais de construção ou monitoramento por vídeo, dependem fortemente da segmentação de imagens (MINAEE et al., 2020). Além disso, carros sem condutor e Sistemas Avançados de Assistência ao Condutor ou *Advanced Driver Assistance* (ADAS), precisam usar a detecção de pedestres ou detectar superfícies navegáveis.

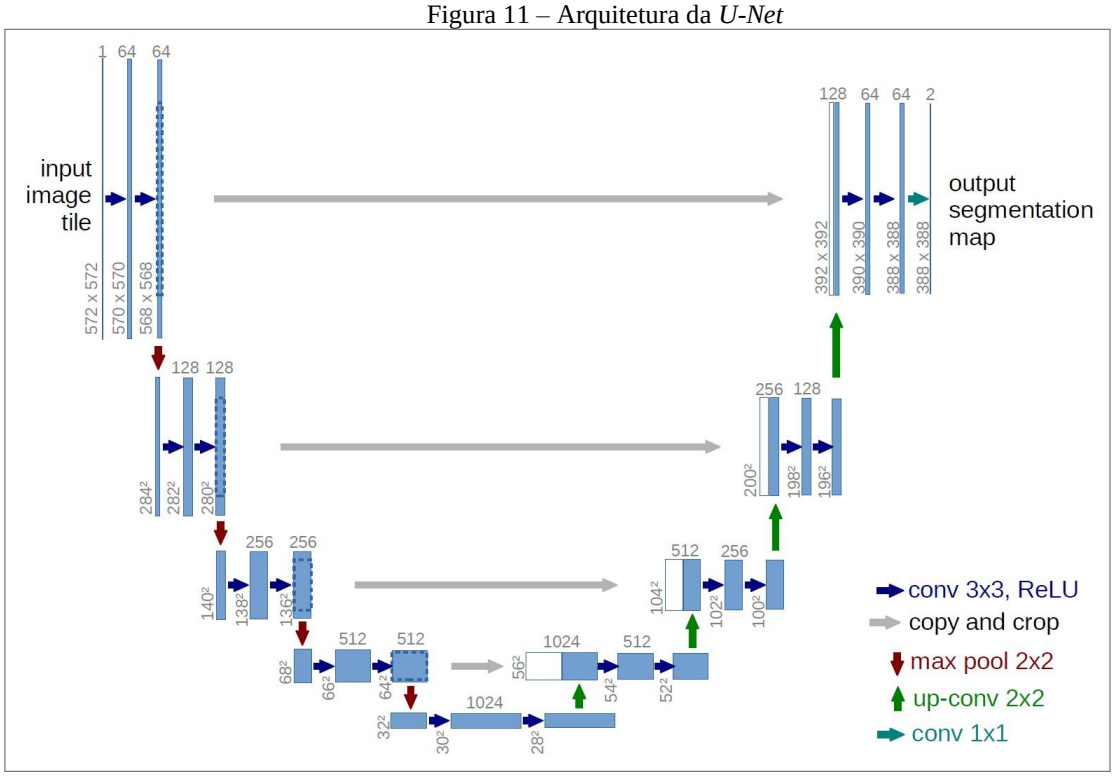
Além disso, a segmentação de imagem é frequentemente usada em ambientes médicos para estimar volumes de tecidos ou extrair limites de tumores. A segmentação dos tumores hepáticos e do fígado, dos tumores cerebrais do cérebro, do disco ótico, da célula, do pulmão, dos nódulos pulmonares e da imagem cardíaca são trabalhos comuns de segmentação de imagem médica (WANG et al., 2020). Dentre as redes de segmentação de imagens existentes, este trabalho dá ênfase a *U-Net* por utilizá-la na fase experimental, e, portanto, ela será detalhada a seguir.

2.6.1 *U-Net*

A rede neural convolucional conhecida como *U-Net* foi criada para a segmentação de imagens biológicas. A rede é construída sobre uma rede totalmente convolucional, porém teve sua arquitetura expandida e alterada para trabalhar com menos imagens de treinamento e produzir uma segmentação mais precisa. A *U-Net* foi sugerida pela primeira vez em um estudo que foi lançado em 2015.

Nesse estudo (RONNEBERGER; FISCHER; BROX, 2015), a ideia principal é aumentar uma rede típica de contração através de camadas sucessivas em que os operadores de *pooling* são substituídos por operadores de *upsampling*. Como resultado, a resolução da saída é aumentada por essas camadas. A fim de obter uma localização precisa dos *pixels* a serem segmentados, as características de alta resolução do caminho de contração da rede são combinadas com a saída em que foi realizado *upsampling*. Uma camada de convolução sucessiva pode aprender a montar uma saída mais precisa com base nessas informações.

Figura 11 – Arquitetura da *U-Net*



Fonte: (RONNEBERGER; FISCHER; BROX, 2015)

A arquitetura da rede é ilustrada na Figura 11. Ela consiste em um caminho de contração (lado esquerdo) e um caminho expansivo (lado direito). A trajetória de contração segue a arquitetura típica de uma rede convolucional e consiste na aplicação repetida de duas convoluções 3 por 3 (convoluções sem *padding*), cada uma seguida por uma unidade linear retificada e uma operação de *max pooling* 2 por 2 com salto 2 para *downsampling*. O número de canais de características é dobrado em cada etapa de *downsampling*. Cada fase do caminho expansivo consiste em um *upsampling* do mapa de características seguido por uma convolução 2 por 2 (*"up-convolution"*), que reduz pela metade o número de canais de características, uma concatenação com o mapa correspondente cortado do caminho de contração e duas convoluções 3 por 3, cada uma seguida por uma ReLU. O recorte do mapa de características é necessário

devido à perda de *pixels* de borda em cada convolução. Na camada final, uma convolução 1 por 1 é usada para mapear cada vetor de característica de 64 componentes para o número desejado de classes. No total, a rede tem 23 camadas convolutivas (RONNEBERGER; FISCHER; BROX, 2015).

3 METODOLOGIA

A metodologia proposta, os dados utilizados, as métricas para avaliação e a composição dos experimentos são discutidos em detalhes nesta seção.

3.1 CARACTERIZAÇÃO DA PESQUISA

Este trabalho pode ser caracterizado quanto ao aspecto metodológico em três componentes: abordagem empregada, objetivos e meios utilizados.

Quanto à abordagem, esta pesquisa caracteriza-se como quantitativa, tendo em vista seu foco na medição objetiva e quantificação de resultados que, por sua vez, foram planejados a fim de alcançar precisão e evitar distorções na etapa de análise e interpretação dos resultados visando a garantia de uma margem de segurança em relação às inferências obtidas pelo modelo (GOMES ALEX SANDRO; GOMES, 2020).

Quanto aos objetivos, esta pesquisa é caracterizada como exploratória, visto o intuito fundamental de descrever o fenômeno que, nesse caso, configura-se como o comportamento preditivo do modelo, de modo que as observações possam ser base para recomendações gerais, sem a formalização de hipóteses de investigação (WAZLAWICK, 2014).

Quanto aos meios utilizados, isto é, aos procedimentos técnicos, esta pesquisa classifica-se como experimental, dado o controle de algumas variáveis experimentais para posterior análise de determinadas variáveis de observação, a fim de investigar se elas podem ser explicadas pelas variáveis experimentais (WAZLAWICK, 2014). Um aspecto predominante nesse tipo de pesquisa é a intervenção sistemática, uma vez que cada intervenção é planejada sistematicamente para que se permita observar os efeitos nas variáveis observadas.

3.2 PROCEDIMENTOS METODOLÓGICOS

Na primeira etapa da pesquisa foi realizada uma investigação bibliográfica para tecer a fundamentação teórica necessária ao desenvolvimento do projeto, obtendo o embasamento científico necessário para sua realização e identificando as perguntas de pesquisa a serem respondidas.

Em seguida, foi realizada uma pesquisa para obter informações sobre bases de dados abertas disponíveis cuja natureza fosse de interesse da pesquisa, sendo feita uma seleção das fontes de dados a serem utilizadas. Tendo acesso aos dados foi realizada a etapa de exclusão de

duplicatas, remoção de dados com qualidade ou resolução inferior a desejada e pré-processamento deles, para padronizar a entrada dos modelos a serem desenvolvidos.

Após isso, foi feita uma nova análise da literatura para identificar quais técnicas comumente utilizadas seriam escolhidas para aplicação neste estudo, juntamente com a escolha dos modelos classificadores que seriam utilizados. Com essa etapa definida, foi realizado a montagem dos experimentos e execução deles, gerando resultados para análise posterior. Os resultados dos experimentos foram, portanto, comparados com outras técnicas de aprendizagem profunda utilizadas na literatura, sendo analisada suas contribuições para a mesma.

3.3 BASES DE DADOS

Foram utilizadas oito bases de dados abertas para o desenvolvimento desta pesquisa, as quais foram separadas em dois *datasets* distintos para realização dos experimentos. Tendo as versões mais recentes das bases coletadas para a aplicação em dezembro de 2021. No total, a base de dados completa disponibilizada contém 22.273 casos de RX classificados como normais, 10.463 casos de pneumonia e 5.050 casos de COVID positivo.

As bases de dados utilizadas foram a *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*, que para fins de praticidade será chamada de OCT (KERMANY, 2018); a *RSNA Pneumonia Detection Challenge* (RSNA) (RSNA, 2018); a *BIMCV-COVID19+: a large annotated dataset of RX and CT images of COVID19 patients* (BIMCV) (VAYÁ et al., 2020); o conjunto de dados *covid-chestxray-dataset*, coletado por Joseph Cohen, o qual será mencionado por Cohen (COHEN et al., 2020); o *Figure 1 COVID-19 Chest X-ray Dataset Initiative* que será chamado apenas de *Figure-1* (CHUNG et al., 2020b); o *Actualmed COVID-19 Chest X-ray Dataset Initiative* referido por *Actual-Med* (CHUNG et al., 2020a); a *RSNA International COVID-19 Open Radiology Database* (RICORD) (RSNA, 2020) e, por fim, *COVID-19 Radiography Database*, uma base de dados aberta proveniente do *site Kaggle*; portanto, será chamada pelo nome do seu *site* de origem (RAHMAN; CHOWDHURY; KHANDAKAR, 2021).

A tabela 1 explana a quantidade de imagens referente a cada classe para todas as bases de dados utilizadas. A escolha de bases de dados abertas foi feita a fim de garantir a reprodutibilidade da pesquisa e para não se restringir a dados regionais que podem ter características discriminantes específicas. Desse modo se buscou a maior variabilidade de dados possível, especialmente para a imagens de raio-X indicando casos positivos da COVID-19.

O primeiro *dataset* criado, o qual chamaremos de *dataset 1*, foi feito de maneira balanceada, realizando *subsampling* das classes para que cada uma tenha um número similar de imagens. As bases utilizadas foram Cohen, *Figure 1*, *Actualmed*, o conjunto de dados do *Kaggle* para Covid-19, RICORD e RSNA, totalizando 7.957 imagens de RX torácico. Para melhor avaliar a capacidade de generalização dos modelos criados, além do conjunto de testes convencionais, são utilizadas as bases restantes (BIMCV e OCT) para criar um conjunto de testes novos com bases que o modelo nunca teve contato.

Dessa maneira, avaliando quanto o modelo realmente diferencia entre os pulmões ou entre outros detalhes das imagens RX, como ruído, anotações, manchas específicas que aparecem apenas em um dos conjuntos etc. A divisão detalhada de treino, teste e validação do *dataset 1* é exposta na tabela 2, enquanto a tabela 3 mostra a quantidade total de imagens para cada classe.

Tabela 1 – Bases de dados utilizadas

Base de Dados	COVID-19	Pneumonia	Normal	Total
OCT	-	4273	1583	5856
BIMCV COVID-19	2470	-	-	2470
Cohen	454	178	18	650
Figure-1	35	-	-	35
Actual-Med	51	-	-	51
Kaggle Covid-19	944	-	-	944
RSNA	-	6012	20672	26684
RICORD	1096	-	-	1096
Total	5050	10463	22273	37786

Fonte: Própria

Tabela 2 – Divisão detalhada de treino, validação e teste para *dataset 1*

Dataset 1										
Base de dados	Treino			Validação			Teste			Total
	COVID-19	Pneumonia	Normal	COVID-19	Pneumonia	Normal	COVID-19	Pneumonia	Normal	
Cohen	291	112	13	72	27	3	80	39	2	639
Figure-1	22	0	0	5	0	0	7			34

Actual-Med	37	0	0	9	0	0	5			51
Kaggle Covid-19	612	0	0	152	0	0	180			944
RICORD	704	0	0	175	0	0	210			1089
RSNA	0	1663	1676	0	415	419		522	505	5200

Fonte: própria

Tabela 3 – Divisão de treino, validação e teste para *dataset 1*

Dataset 1				
Classe	Treino	Validação	Teste	Total
COVID-19	1666	413	482	2561
Pneumonia	1775	442	561	2778
Normal	1689	422	507	2618

Fonte: própria

No segundo *dataset* criado, o qual chamaremos de *dataset 2*, também houve *subsampling* das classes, no entanto se manteve um desbalanceamento para a classe de pulmões considerados saudáveis. As bases utilizadas foram Cohen, *Figure 1*, *Actualmed*, o conjunto de dados do *Kaggle* para Covid-19, BIMCV e OCT, totalizando em 9.952 imagens de raio-X torácico para o conjunto. Para melhor avaliar a capacidade de generalização dos modelos criados, também foi elaborado um segundo conjunto de testes secundáricom as bases restantes (RICORD e RSNA), ambos conjuntos de testes secundários são detalhados na tabela 6. A divisão detalhada de treino, teste e validação do *dataset 2* é exposta na tabela 4, enquanto a tabela 5 mostra a quantidade total de imagens para cada classe.

Tabela 4 – Divisão detalhada de treino, validação e teste para *dataset* 2

Dataset 2										
Base de dados	Treino			Validação			Teste			Total
	COVID-19	Pneumonia	Normal	COVID-19	Pneumonia	Normal	COVID-19	Pneumonia	Normal	
Cohen	283	112	11	70	27	2	101	39	5	650
Figure-1	20	-	-	5	-	-	10	-	-	35
Actual-Med	33	-	-	8	-	-	10	-	-	51

Kaggle Covid-19	616	-	-	153	-	-	175	-	-	944
OCT	-	2764	1018	-	691	254	-	764	311	5802
BIMCV COVID-19	1614	-	-	403	-	-	453	-	-	2470

Fonte: própria

Tabela 5 – Divisão de treino, validação e teste para *dataset 2*

Dataset 2				
Classe	Treino	Validação	Teste	Total
COVID-19	2566	639	749	3954
Pneumonia	2876	718	803	4397
Normal	1029	256	316	1601

Fonte**:** Própria

Tabela 6 – Conjuntos de testes secundários

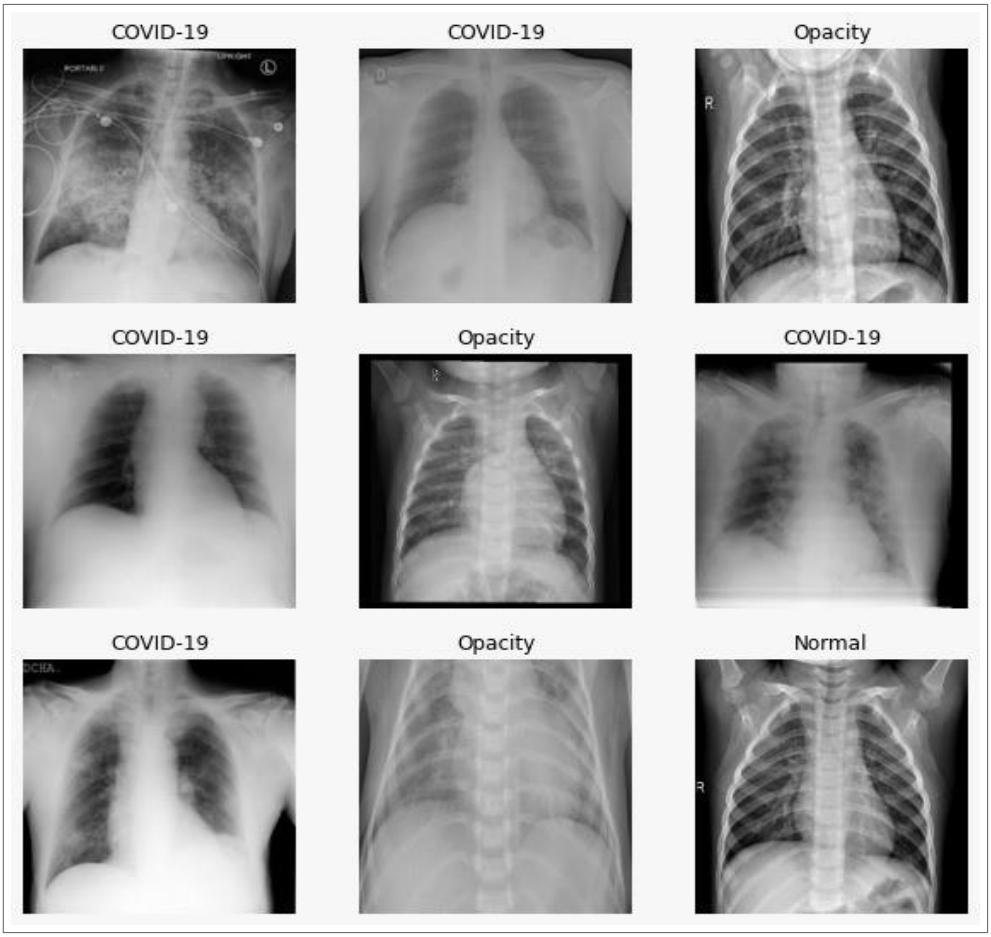
Conjunto de testes secundário 1			
Base de dados	COVID-19	Pneumonia	Normal
OCT	-	1228	468
BIMCV	723	-	-
Conjunto de tests secundário 2			
RSNA	-	789	772
RICORD	322	-	-

Fonte**:** Própria

A escolha de treinar os modelos com 2 conjuntos distintos foi feita para poder realizar a comparação entre a capacidade de generalização de um mesmo modelo treinado com bases diferentes, de modo que analisar se os conjuntos de dados hoje disponíveis juntamente com as técnicas mais utilizadas são suficientes para obter bons resultados em aplicações reais. Portanto, se justifica também a necessidade de ter um conjunto de testes secundário em que outros dados de bases diferentes nunca vistas pelos modelos são utilizadas para visualizar como eles se comportariam em situações de aplicação real, em que as imagens de RX utilizadas possuem variações e características diferentes das que foram usadas para o treinamento.

Ainda sobre as imagens RX coletadas, elas têm diferentes tonalidades de cinza, dimensões e características, sendo transformadas para escala de cinza e redimensionadas para as dimensões de 300 por 300 pixels como forma de pré-processamento de dados. A figura 12 ilustra amostras de imagens de raios-X de pacientes infectados com COVID-19, de pacientes considerados normais e de pacientes com pneumonia não proveniente da COVID-19 (denominado por *opacity*).

Figura 12 – Amostras de *Raio-X*



Fonte: (VAYÁ et al., 2020; KERMANY, 2018)

3.4 MÉTRICAS

3.4.1 Acurácia e erro

Para tarefas de classificação, muitas vezes é medida a acurácia do modelo. Acurácia é a proporção de exemplos para os quais o modelo produz o resultado correto.

$$Acurácia = \frac{\text{número de predições corretas}}{\text{número total de predições}} \tag{3.1}$$

Informações equivalentes às dadas pela acurácia podem ser obtidas ao medir a taxa de erro, isto é, a proporção de exemplos para os quais o modelo produz uma saída incorreta. Muitas vezes essa taxa é referida por taxa de erro, em que a perda esperada é uma medida que varia de 0 a 1. A perda em um determinado exemplo é 0 se estiver corretamente classificada e 1 se não estiver (HAYKIN, 2008).

A acurácia só se torna uma medida confiável se houver um número igual ou semelhante de amostras pertencentes a cada classe. Por exemplo, considerando que há 95% de amostras da classe A e 5% de amostras da classe B em um conjunto de treinamento, o modelo pode facilmente obter 95% de acurácia simplesmente prevendo todas as amostras de treinamento como pertencentes à classe A. No entanto, quando o mesmo modelo é testado em um conjunto de teste com 40% de amostras da classe A e 60% de amostras da classe B, a precisão do teste cairia para 40%. A acurácia pode ser uma boa métrica, mas também pode causar a falsa sensação de alcançar um alto desempenho.

3.4.2 Matriz de confusão

A matriz de confusão é uma medida bastante popular utilizada na solução de problemas de classificação. Ela pode ser aplicada à classificação binária, bem como para problemas de classificação multiclasse. Ela é definida como a tabela que descreve o desempenho de um modelo de classificação em um conjunto de dados de teste para os quais os valores verdadeiros são conhecidos.

As matrizes de confusão representam as contagens dos valores previstos e reais dos rótulos dos dados. A saída “TN” significa verdadeiro negativo, que mostra o número de exemplos negativos classificados corretamente. Da mesma forma, “TP” significa verdadeiro positivo, que indica o número de exemplos positivos classificados com exatidão. O termo “FP” mostra o valor falso positivo, ou seja, o número de exemplos reais negativos classificados como positivos; o “FN” significa um valor falso negativo que é o número de exemplos reais positivos classificados como negativos (BAJAJ; SINHA, 2022).

3.4.3 Precisão

Precisão ou *precision* como uma métrica de avaliação tenta responder à seguinte pergunta: Que proporção de identificações positivas foi realmente correta?

Quadro 3 – Estrutura de uma matriz de confusão para classificação binária

		Valor Predito	
		Sim	Não
Valor Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte**:** Própria

A precisão explica quantos dos casos previstos corretamente se revelaram realmente positivos (MURPHY, 2022). É útil nos casos em que as falsas positivas são uma preocupação maior do que os falsos negativos. A precisão pode ser representada pela equação abaixo:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

3.4.4 Recall

O *Recall* mostra a proporção de previsões positivas corretas dentre todos os pontos positivos que um modelo poderia ter feito. Para calculá-la, é necessário dividir todos os verdadeiros positivos pela soma de todos os verdadeiros positivos e falsos negativos no conjunto de dados (POWERS, 2020). Dessa forma, o *recall* fornece uma indicação de previsões positivas perdidas, ao contrário da métrica de precisão que foi explicada acima. A recordação pode ser representada pela equação abaixo:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

3.4.5 F1-Score

Comparar modelos com baixa precisão e alto *recall* (ou vice-versa) pode ser uma tarefa complexa e custosa. O *F1-score*, portanto, existe de tal maneira que facilita essa comparação, ajudando a medir o *recall* e a precisão ao mesmo tempo. Ele emprega a média harmônica das duas métricas no lugar da média aritmética, penalizando mais os valores extremos (MURPHY, 2022). O *F1-Score* pode ser representado pela seguinte equação:

$$Recall = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3.4}$$

O *F1-Score* é geralmente utilizado quando existem dados desequilibrados entre as classes. Na maioria das aplicações da classificação em dados reais, encontra-se uma distribuição de classe desequilibrada e, portanto, o *F1-Score* é uma métrica considerada melhor para avaliar um modelo do que a precisão.

No entanto, ainda se opta pela utilização da precisão em estudos devido ao *F1-Score* ser menos interpretável. Precisão e *recall* são métricas mais interpretáveis pela precisão em medir o erro tipo 1 e o *recall* medir o erro tipo 2. Sendo o *F1-Score* o *trade-off* entre estes dois.

3.4.6 Curva ROC e AUC

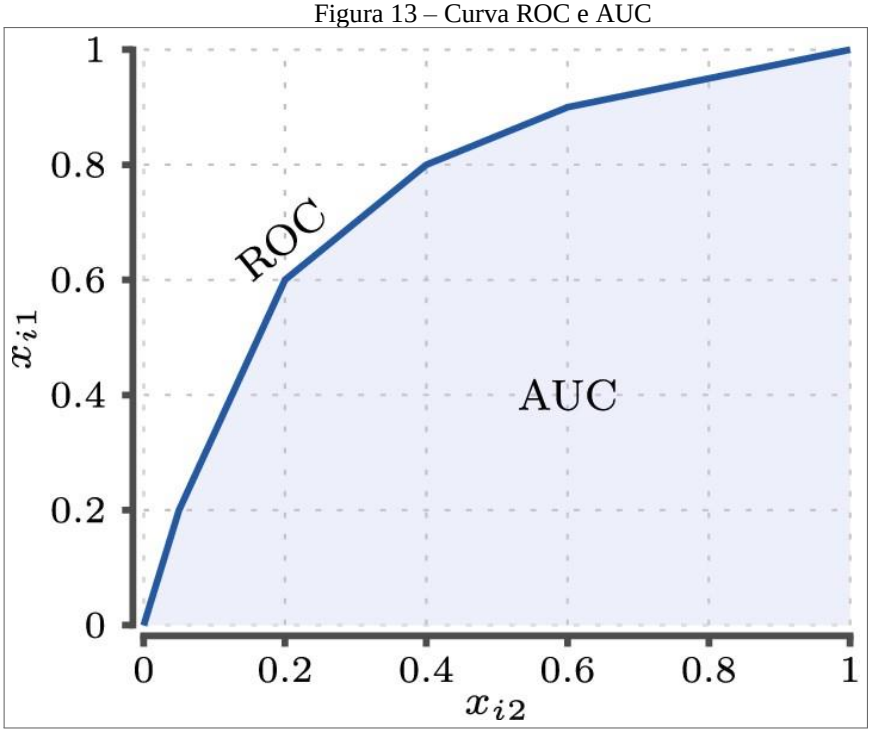
A Característica do Operador Receptor ou *Receptor Operator Characteristic* (ROC) é uma curva de probabilidade que traça a TPR (*True Positive Rate*, também conhecida como sensibilidade) contra a FPR (*False Positive Rate*) em vários valores limiares (FAWCETT, 2006). A curva ROC é traçada com TPR em relação à FPR, na qual TPR está no eixo y e FPR está no eixo x. As equações para TPR e FPR são representadas a seguir:

$$TPR = \frac{TP}{TP + FN} \tag{3.5}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.6}$$

A Área Sob a Curva, *Area under the Curve* (AUC), é a medida da capacidade de um classificador de distinguir entre as classes. Utilizando a figura 13, verfica-se que a AUC corresponde a área abaixo da curva ROC entre os eixos X e Y.

Figura 13 – Curva ROC e AUC



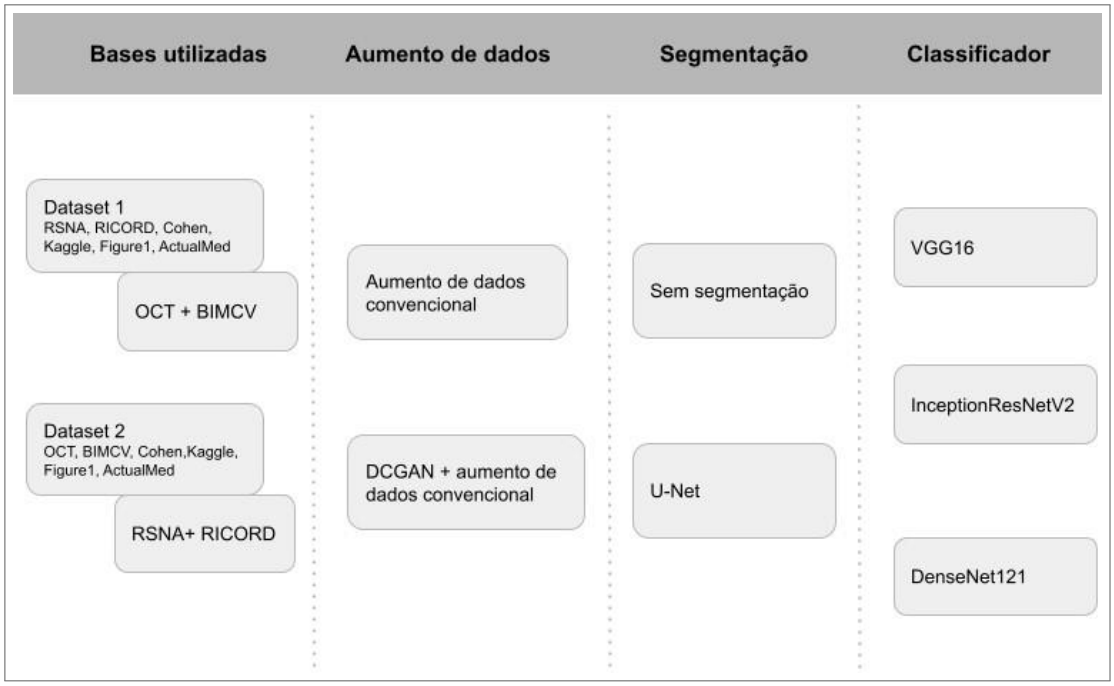
Fonte: (TATTI, 2021)

AUC e Curva ROC são medidas de desempenho para os problemas de classificação em várias configurações de *thresholds*. A ROC é uma curva de probabilidade, e AUC representa o grau ou medida de separabilidade. Ela informa quanto o modelo é capaz de distinguir entre as classes. Assim, quanto maior a AUC, melhor o modelo está prevendo a classe 0 como 0 e a classe 1 como 1.

3.5 EXPERIMENTOS

Os experimentos deste trabalho partiram de três técnicas comumente empregadas na literatura para detecção de COVID-19, em que imagens radiológicas foram combinadas e avaliadas, a fim de tentar construir uma *pipeline* com ótimo desempenho na classificação multiclasse entre pulmões com pneumonia causada por COVID-19, pulmões normais e pulmões apresentando pneumonia não proveniente da COVID-19. Além disso, busca-se avaliar como os modelos criados poderiam se comportar com dados reais e visualizar qual combinação de técnicas alcança a melhor performance para esse cenário. A figura 14 exemplifica todas possíveis etapas de uma *pipeline*. Em seguida, essas serão explicadas detalhadamente.

Figura 14 – Etapas da *pipeline*



Fonte: Própria

Foram determinadas quatro etapas para as *pipelines* criadas, a etapa de escolha da base de dados, do método de aumento de dados, da atividade de segmentação ou não dos dados e, por fim, do classificador a ser utilizado. Assim, têm-se quatro possíveis combinações de *pipelines*: a *baseline*, a qual não há aumento de dados e tão pouco segmentação, apenas os conjuntos e classificadores deles são variados; a *pipeline* utilizando DCGAN, a qual não há segmentação, uma vez que o aumento de dados se dá pela criação de imagens sintéticas e pela utilização do aumento convencional; a *pipeline* com segmentação, em que a U-Net é utilizada para segmentar as imagens provenientes das bases de dados; por fim, a *pipeline* completa que combina as técnicas de segmentação com a geração de imagens sintéticas pela DCGAN.

Dessa forma, foram escolhidas as técnicas de *transfer learning*, aumento de dados e segmentação de dados para compor *pipelines* de treinamento de modelos. Cada *pipeline* foi repetida 20 vezes com separações distintas entre teste, treino e validação, para aleatorizar esses *datasets* sem enviesar o treinamento. Por exemplo, utilizando apenas um conjunto em que aquela separação de dados específica resultou em uma acurácia de teste muito alta. Ao realizar 20 experimentos com separações de treino, teste e validação diferentes, foi possível obter as médias e desvios de padrões das métricas de cada *pipeline* e ter uma ideia mais assertiva de como cada modelo se comportaria na vida real.

Na etapa de *transfer learning* foram escolhidas três CNNs: a *DenseNet-121*, a VGG-16 e a *InceptionResNet-V2*. Todas as CNNs utilizadas estão entre as melhores performances na

classificação para o *dataset* da *ImageNet* e têm arquiteturas bastante diferentes entre si, com vantagens e desvantagens em suas estruturas.

As *DenseNets* têm diversas vantagens atrativas: aliviam o problema do "gradiente de fuga"; fortalecem a propagação de características; incentivam a reutilização de características; reduzem substancialmente o número de parâmetros de treinamento (HUANG et al., 2016). A *DenseNet-121* apresenta acurácia Top-1 no *ImageNet* de 75,0%, acurácia Top-5 de 92,3%, 8,1 milhões de parâmetros e profundidade de 242 camadas.

Já a rede VGG-16 foi escolhida para se ter uma rede clássica como *benchmark*, ou seja, um ponto de referência contra o qual as outras redes podem ser comparadas ou avaliadas. Em relação ao conjunto de dados da *ImageNet,* ela tem uma performance inferior à densenet, obtendo 71,3% de acurácia para Top-1 e 90,1% para Top-5. Essa CNN tem 138,4 milhões de parâmetros e 16 camadas de profundidade.

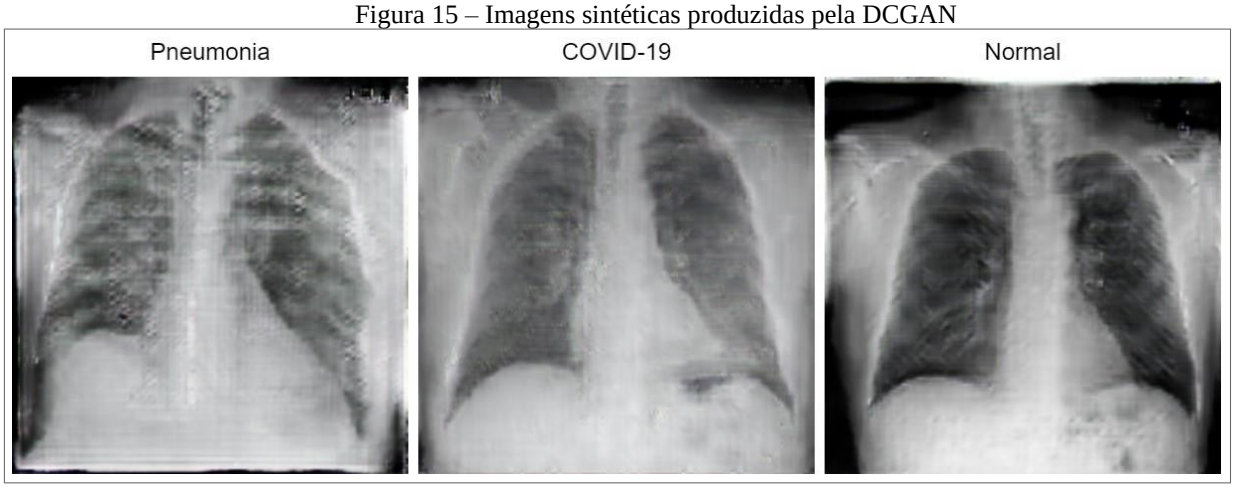
A *InceptionResNet-V2* tem como vantagens seu alto desempenho no *ImageNet* com baixo custo computacional comparada a outras redes de desempenho semelhante, redução do problema de degradação e tempo de treinamento reduzido. Dentre as redes escolhidas, ela é a que comporta as melhores acurácias para o *ImageNet*, chegando a 80,3% para o Top-1 e 95,3% para o Top-5. Com 449 camadas, essa CNN tem 55,9 milhões de parâmetros.

Para etapa de aumento de dados, foi utilizada uma aplicação convencional, isto é, a realização de pequenas mutações nas imagens originais com objetivo de aumentar a variabilidade do conjunto de dados. As mutações escolhidas foram o espelhamento da imagem, pequenas deslocações, mudanças de escala e rotações em conjunto, transformações elásticas, brilho aleatório, contraste aleatório e aplicação de *gamma* aleatório, sendo todas mutações escolhidas para simular condições diferentes de obtenção de imagens RX. Essa etapa é aplicada em todos os pipelines existentes, uma vez que esses processos têm sido constantemente utilizados para melhorar o desempenho de modelos de aprendizagem de máquina (CHLAP et al., 2021).

Opcionalmente também foi utilizada uma DCGAN treinada com 300 épocas no conjunto de treino selecionado para a *pipeline,* com o objetivo de criar imagens sintéticas e adicioná-las ao conjunto de treino já anteriormente separado. Assim, enriquece a variabilidade dos dados, possibilitando a comparação entre o uso ou não de dados sintéticos como técnica de aumento de dados. No total 1500 imagens foram adicionadas a cada conjunto de treino para as *pipelines* que utilizaram essa técnica, sendo 500 imagens pertencentes a cada classe. A escolha do número de épocas justifica-se devido ao fato de que a partir do número 300 os resultados convergiram, e não houve mais diferença significativa entre eles. A escolha da DCGAN, para

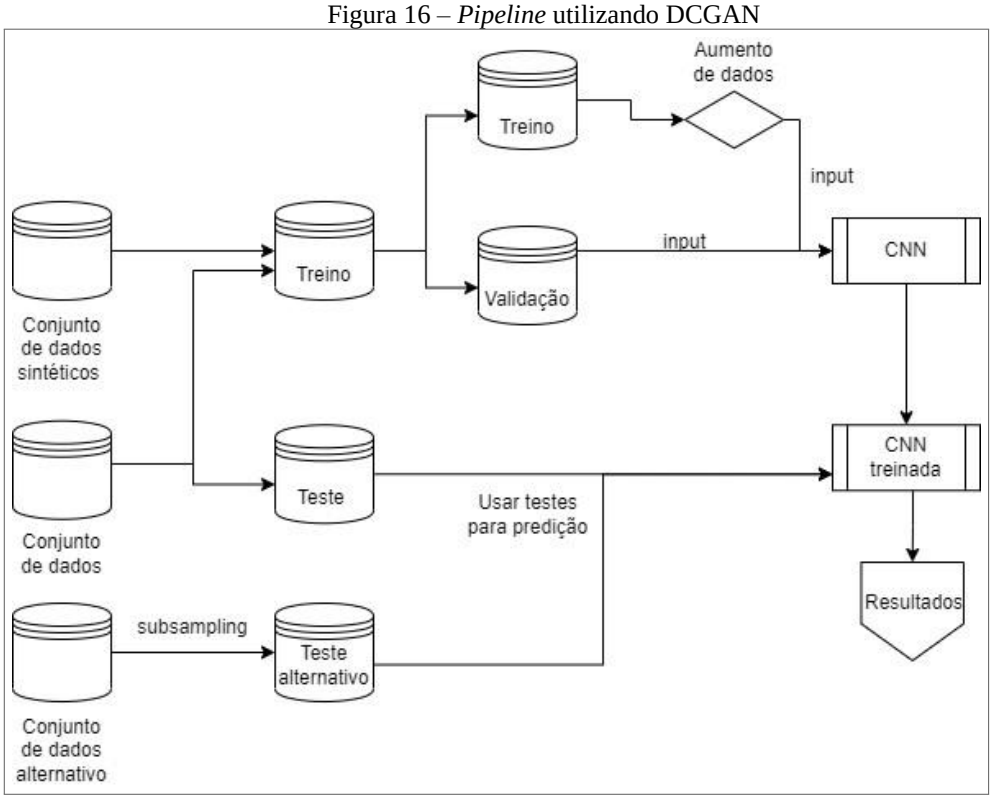
esta tarefa, foi baseada em seu bom desempenho no estudo de Cavalcanti e Berton (2021). Na figura 15 tem um exemplo de imagens geradas pela DCGAN. A figura 16 exemplifica a *pipeline* que utiliza a DCGAN como técnica extra de aumento de dados.

Figura 15 – Imagens sintéticas produzidas pela DCGAN



Fonte: própria

Figura 16 – *Pipeline* utilizando DCGAN



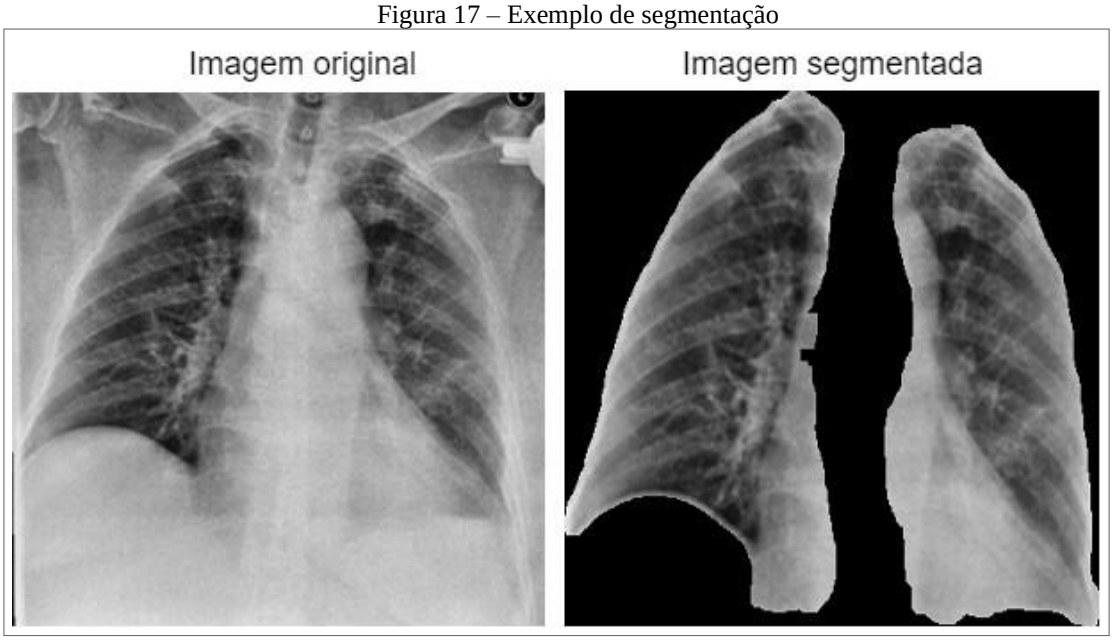
Fonte: própria

A *U-net* foi usada na etapa de segmentação, na medida em que é uma técnica de segmentação de imagem desenvolvida, sobretudo para análise de imagens médicas, podendo

segmentar com precisão imagens usando uma quantidade escassa de dados de treinamento. Ela apresenta uma adoção extensiva como a principal ferramenta para tarefas de segmentação em imagens médicas (SIDDIQUE et al., 2021). Essa técnica também foi utilizada por Teixeira et al. (2021) na aplicação de segmentação de RX, a fim auxiliar na classificação do COVID-19.

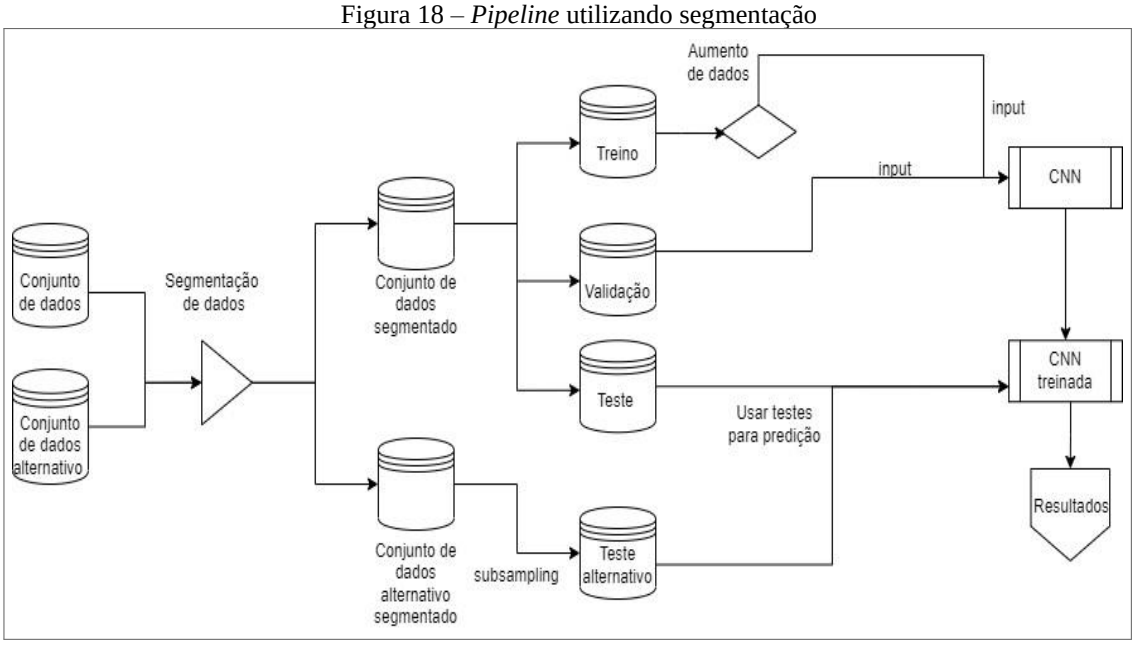
Dessarte, foi seguida a implementação da *U-Net* proposta por Teixeira et al. (2021) com algumas modificações, pois não foram feitas máscaras manuais, apenas os conjuntos de dados JSRT (JSRT, 1998), Shenzen, Montgomery (JAEGER et al., 2014) e Cohen (2020). Foram utilizados com 385, 566, 138 e 508 imagens cada, respectivamente, totalizando 1597 imagens na base unificada. O modelo foi treinado com 100 épocas e aplicado aos conjuntos coletados anteriormente. A escolha do número supracitado de épocas justifica-se na convergência observada do modelo. A figura 17 demonstra um exemplo de segmentação realizada pelo modelo, enquanto a figura 18 mostra a *pipeline* com a segmentação incorporada.

Figura 17 – Exemplo de segmentação



Fonte: própria

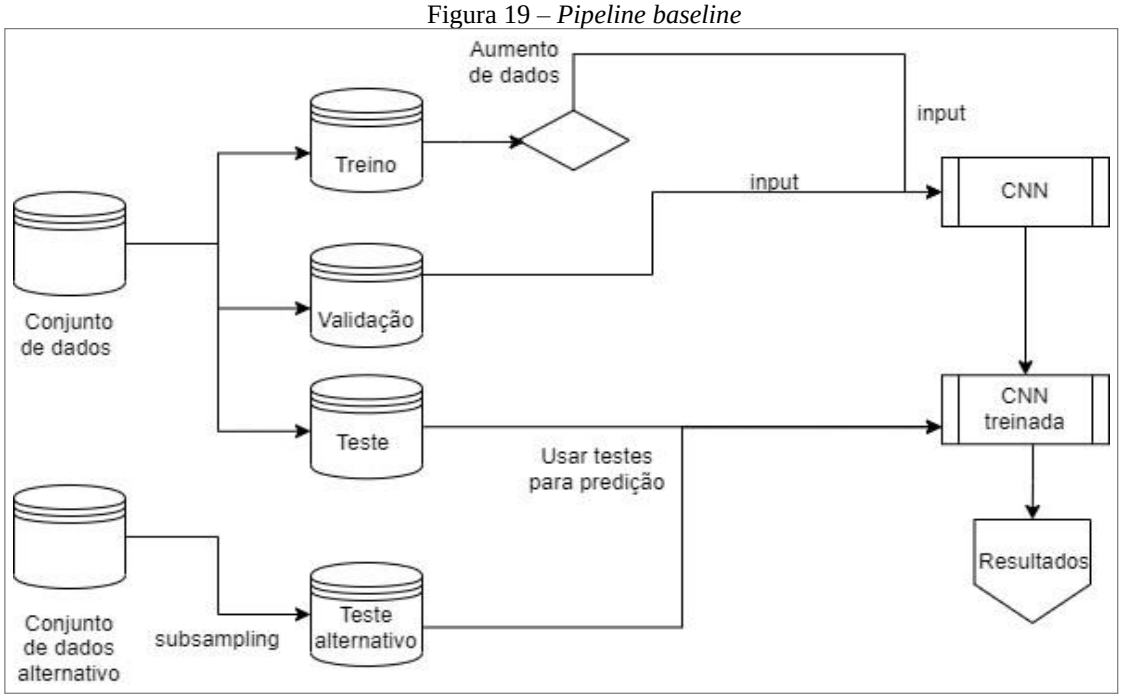
Figura 18 – *Pipeline* utilizando segmentação



Fonte: própria

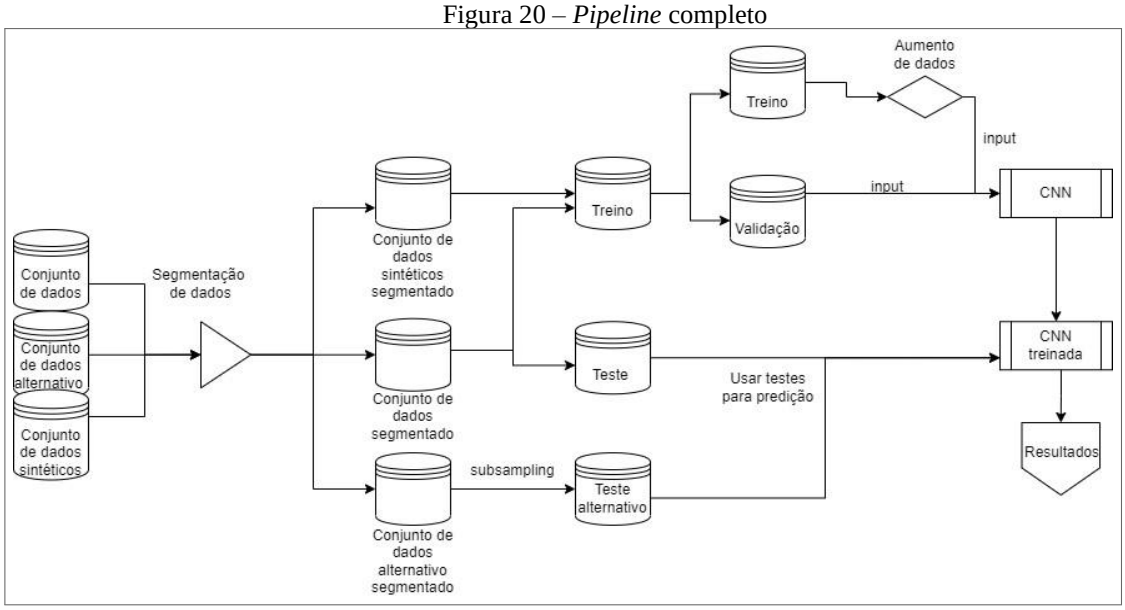
Foram criadas 4 combinações possíveis para cada modelo de *transfer learning*, sendo elas: apenas aumento de dados convencional com segmentação, apenas aumento de dados convencional sem segmentação, DCGAN com segmentação e DCGAN sem segmentação. Totalizando 12 experimentos para o *dataset* 1 e 3 experimentos para o *dataset* 2. A figura 19 mostra o *pipeline* sem segmentação ou aumento de dados com GAN, ou seja, o pipeline *baseline*, enquanto a figura 20 exemplifica o *pipeline* completo.

Figura 19 – *Pipeline baseline*



Fonte: própria

Figura 20 – *Pipeline* completo



Fonte: própria

Os modelos retreinados por meio de *transfer learning* foram inicializados com os pesos da *ImageNet e* treinados com 200 épocas cada. Em 50 épocas, apenas camadas finais foram submetidas ao retreinamento; em 150 épocas, no entanto, foi realizado o ajuste fino, exceto pela VGG-16, que tem 100 épocas para ajuste fino, visto que o tempo de treinamento em experimentos preliminares excedia o tempo das outras redes quando era utilizada 200 épocas sem aumento de performance evidente.Todas as CNNs foram treinadas valendo-se do algoritmo de Adam como otimizador, com taxa de aprendizagem inicial de 0,0001. Foi operado um *checkpointer* para salvar o melhor modelo e seus pesos. Assim, a taxa de aprendizagem foi dinamicamente reduzida ao longo do treinamento *finetuning* ao aplicar a função *ReduceLROnPlateau* da biblioteca *keras*. Apenas a *InceptionResNetV2* teve todas suas camadas retreinadas no *finetuning*, uma vez que em experimentos preliminares, notou-se a obtenção de resultados melhores com pouca penalização em seu tempo de treinamento. A escolha do número de épocas se justifica na convergência observada dos modelos.

Tabela 7 – Configurações das CNNs escolhidas

CNN	Número de camadas	Camadas treináveis em finetuning	épocas finetuning	épocas totais	taxa de aprendizagem inicial
VGG16	19	9	100	150	0,0001
DenseNet121	427	277	150	200	0,0001

InceptionResNetV2	780	780	150	200	0,0001

Fonte: Própria

Os experimentos foram executados utilizando tensorflow-gpu 2.6.0 e keras 2.6.0 na versão 3.8.11 do python, em uma máquina com o processador intel core i7 da décima geração, com 16GB de memória RAM e placa de vídeo dedicada GeForce RTX 2070 com 8GB de memória. A versão da CUDA foi a 11.4, o código referente aos experimentos pode ser encontrado no seguinte repositório: https://github.com/silveiraluiza/covid-pred/.

4 RESULTADOS E DISCUSSÕES

Com o objetivo de comparar quantitativamente todos os experimentos executados, foram calculadas as acurácias de treino, teste e teste secundário, assim como o *F1-score*, a precisão e o *recall*. Também foram calculadas as curvas ROC do modelo em relação a classe COVID-19 e sua AUC. Todas essas métricas foram calculadas para cada uma das 20 execuções de cada *pipeline*, e as métricas apresentadas aqui são as médias obtidas dessas execuções. A tabela 8 mostra o *F1-score* de teste e o tempo médio de treinamento dos melhores modelos de cada *pipeline*. Os melhores modelos foram escolhidos levando em consideração sua performance no conjunto de testes secundário.

Tabela 8 – Melhores modelos de cada *pipeline*

Pipeline	Base de dados	CNN		F1-Score	Tempo médio de treinamento
Completa	Dataset 1	InceptionResNetV2		0,936	7:50:37
DCGAN	Dataset 1	InceptionResNetV2		0,934	7:50:50
Segmentação	Dataset 1	InceptionResNetV2		0,908	6:20:14
	Dataset 2	DenseNet121		0,957	6:53:58
Baseline	Dataset 1	VGG16		0,931	4:28:40
	Dataset 2	VGG16		0,964	5:37:30

Fonte: Própria

É possível observar que os *F1-Scores* dos modelos estão bastante próximos, tendo o modelo no qual foi aplicado o uso de dados sintéticos e segmentação um *F1-Score* de 0,936, o modelo em que apenas foi aplicado o uso de dados sintéticos sem segmentação 0,934 de *F1-Score* e o modelo *baseline* um *F1-Score* de 0,931. No entanto, observa-se uma leve queda de performance ao comparar o modelo com segmentação e o *baseline*, sendo o *F1-Score* do primeiro de 0,908. Teixeira et al. (2021) observou esse mesmo comportamento em sua pesquisa e constatou que, ao selecionar apenas a área do pulmão para o treinamento, a segmentação remove algumas das características das imagens que poderiam estar enviesando o modelo.

Também verifica-se que o desempenho médio de todos os modelos está próximo dos desempenhos encontrados na literatura, sendo comparáveis diretamente aos trabalhos de Chakraborty; Murali; Mitra (2022) e Nefoussi; Amamra; Amarouche (2021). E, no caso dos modelos em que houve a aplicação de segmentação de imagens, foi obtido uma performance

superior ao modelo de Teixeira et al. (2021), realizando uma aplicação bastante similar ao seu estudo.

A tabela 9 mostra o *F1-score* obtido a partir do *dataset* secundário dos melhores modelos de cada *pipeline.*

Tabela 9 – Melhores modelos de cada *pipeline* para *dataset* secundário

Pipeline	Base de dados	CNN		F1-Score
Completa	Dataset 1	InceptionResNetV2		0,325
DCGAN	Dataset 1	InceptionResNetV2		0,374
Segmentação	Dataset 1	InceptionResNetV2		0,551
	Dataset 2	DenseNet121		0,151
Baseline	Dataset 1	VGG16		0,376
	Dataset 2	VGG16		0,206

Fonte: Própria

Utilizando o teste de Kolmogorov-Smirnov, foi verificado que os conjuntos de *F1-Scores* do teste secundário obtidos com as 20 repetições de execução dos modelos não seguem uma distribuição normal, visto que seu p-valor era menor do que 0,05. Dessa forma se pôde aplicar o teste U de Mann-Whitney para o fim de comparar os resultados de maneira pareada.

Para os modelos do *dataset* 1 o teste de Mann-Whitney foi aplicado para todas as combinações de *pipelines*. Para a comparação entre a *baseline* e DCGAN, as distribuições foram apontadas como iguais, com seu p-valor sendo igual a 0,903. Para a comparação entre as *pipelines* restantes as distribuições foram apontadas como diferentes, tendo seu p-valor menor do que 0,05. A comparação entre os dois modelos do *dataset* 2 resultou em distribuições diferentes, com seu p-valor sendo igual a 0,001.

A seguir os resultados de cada *pipeline* serão explorados com maior profundidade, sendo expostas as suas matrizes de confusão e curvas ROC.

4.1 RESULTADOS *DATASET* 1

4.1.1 *Pipeline baseline*

Tabela 10 – Métricas da pipeline baseline para *dataset 1*

		DenseNet121	InceptionResNetV2	VGG16
Acurácia de teste	média	0,931	0,937	0,931
	desvio padrão	0,006	0,006	0,011
	máximo	0,940	0,950	0,946
Acurácia de teste secundário	média	0,405	0,398	0,422
	desvio padrão	0,033	0,044	0,052
	máximo	0,495	0,477	0,575
Acurácia de treino	média	0,995	0,999	0,991
	desvio padrão	0,013	0,001	0,022
	máximo	0,999	1,000	1,000
F1-Score de teste	média	0,931	0,937	0,931
	desvio padrão	0,006	0,006	0,011
	máximo	0,940	0,950	0,946
F1-Score de teste secundário	média	0,361	0,345	0,376
	desvio padrão	0,044	0,063	0,074
	máximo	0,467	0,453	0,572
F1-Score de treino	média	0,995	0,999	0,991
	desvio padrão	0,013	0,001	0,022
	máximo	0,999	1,000	1,000
Recall de teste	média	0,931	0,937	0,931
	desvio padrão	0,006	0,006	0,011
	máximo	0,940	0,950	0,946
Recall de teste secundário	média	0,405	0,398	0,422
	desvio padrão	0,033	0,044	0,052
	máximo	0,495	0,477	0,575
Recall de treino	média	0,995	0,999	0,991
	desvio padrão	0,013	0,001	0,022
	máximo	0,999	1,000	1,000
Precisão de teste	média	0,931	0,937	0,932
	desvio padrão	0,006	0,006	0,011
	máximo	0,940	0,950	0,946
	média	0,598	0,616	0,655

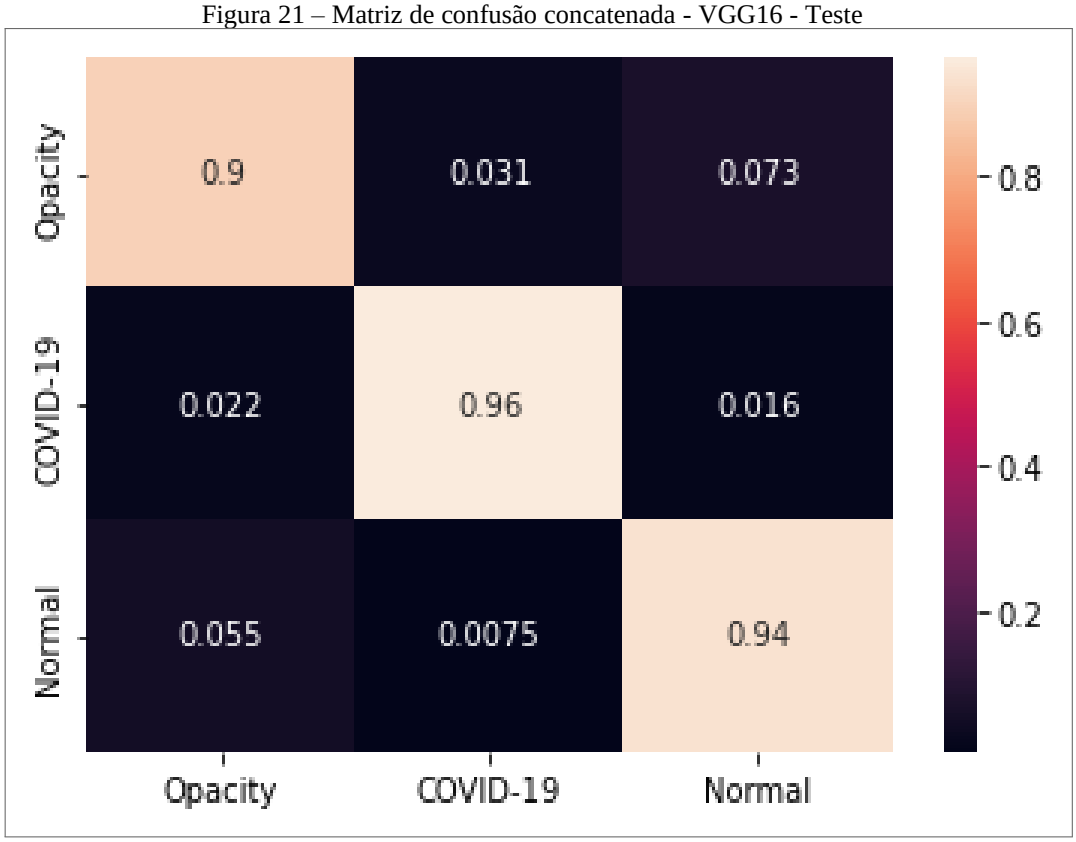
Precisão de teste secundário	desvio padrão	0,024	0,036	0,031
	máximo	0,649	0,681	0,717
	média	0,995	0,999	0,991
Precisão de treino	desvio padrão	0,013	0,001	0,021
	máximo	0,999	1,000	1,000

Fonte: Própria

Na tabela 9 foram expostas as médias, desvios padrões e o desempenho máximo de todas as métricas calculadas para as *pipelines baseline* executadas com o *dataset* 1. Para o *recall*, a precisão e o *F1-score*, utilizou-se a média aritmética ponderada. A partir desses dados, é possível observar que os modelos apresentam resultados bastante similares com os encontrados na literatura, especialmente porque não são expostas as médias dos modelos encontrados, e sim seus melhores desempenhos. Para esse conjunto de experimentos, o melhor desempenho na acurácia e no *F1-score* foi de 95%, superando, assim, os desempenhos encontrados nos trabalhos de Chakraborty; Murali; Mitra (2022), Nefoussi; Amamra; Amarouche (2021), Bhattacharyya et al. (2022), Teixeira et al. (2021) e Nikolau et al. (2021). Ao analisar os desempenhos médios, também é possível os descrever como satisfatórios por estarem próximos aqueles descritos na literatura apresentada anteriormente.

Outro ponto a ser observado é a diferença entre as acurácias e *F1-scores* de teste e as acurácias e *F1-scores* obtidas com a predição na base de testes secundária. Há uma queda pronunciada no desempenho dos modelos quando são aplicados a um conjunto totalmente novo, o que pode implicar uma má capacidade de generalização por parte dos modelos. No trabalho de Teixeira et al. (2021), o melhor *F1-score* obtido foi de 77% em um experimento semelhante; no entanto, nele foram consideradas apenas 2 classes: COVID-19 positivo e negativo. Vale destacar que ambas bases de treinamento utiliza os dados RSNA, enquanto no experimento realizado neste trabalho não existe nenhuma base em comum entre o conjunto de treino e o conjunto de testes secundário. Ainda assim, em seu melhor cenário, foi obtido um *F1-score* de 57,2%. É apresentada a seguir a matriz de confusão concatenada, isto é, que apresenta todos os resultados de todas as 20 execuções do modelo com *folds* diferentes de treino, teste e validação, do melhor modelo *baseline*.

Figura 21 – Matriz de confusão concatenada - VGG16 - Teste

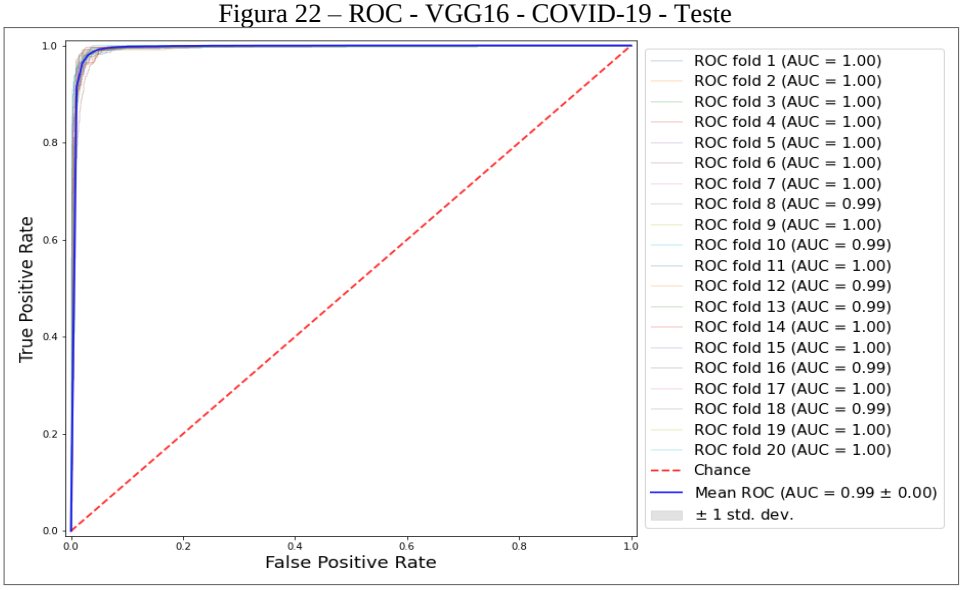


Fonte: Própria

A partir dessa matriz é possível visualizar a classe de imagens que apresentam o pulmão com pneumonia oriunda da COVID-19 com mais verdadeiros positivos. Além disso, cabe ressaltar que a classe de pneumonia está sendo mais confundida com a normal do que com COVID-19, um comportamento não esperado, visto que tanto a pneumonia viral ou bacteriana, como a COVID-19, faz com que padrões de vidro fosco apareçam no pulmão afetado.

No entanto, esse comportamento é justificado porque maior parte dos dados de pneumonia e pulmão normal serem provenientes da mesma base de dados, enquanto os dados de COVID-19 vêm de bases diferentes. Dessa forma, o modelo pode ter aprendido a procurar por características específicas referentes a base predominantemente utilizada para pneumonia e pulmão normal para diferenciá-las da COVID-19. Pode-se visualizar claramente o grau alto de separabilidade da classe COVID-19 na figura 22, que representa a curva ROC e a AUC associadas a classe COVID-19.

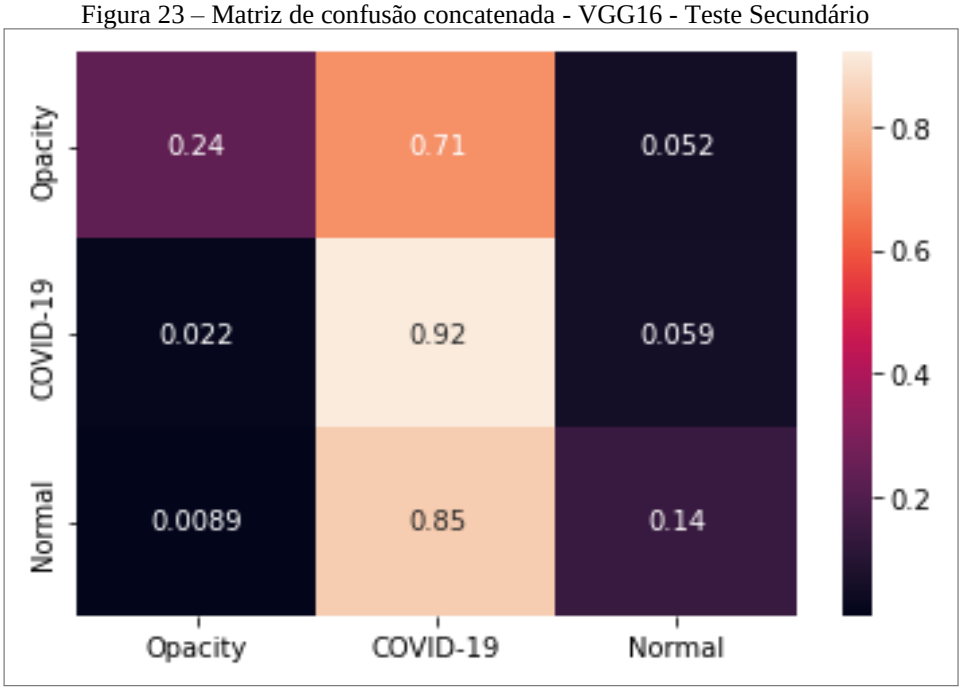
Figura 22 – ROC - VGG16 - COVID-19 - Teste



Fonte: Própria

A figura 23 representa a matriz de confusão concatenada para o conjunto de dados secundários, sendo visível um enviesamento para a classe COVID-19. A maioria das imagens foi classificada apresentando pulmões com pneumonia proveniente da COVID-19. Os pulmões normais foram erroneamente classificados como pulmões com COVID-19 mais do que pulmões os com pneumonia, o que mostra que o modelo ainda consegue distinguir essas duas classes, pneumonia e COVID-19, um pouco melhor do que normal e COVID-19.

Figura 23 – Matriz de confusão concatenada - VGG16 - Teste Secundário



Fonte: Própria

4.1.2 *Pipeline* com segmentação

Tabela 11 – Métricas da *pipeline* com segmentação para *dataset 1*

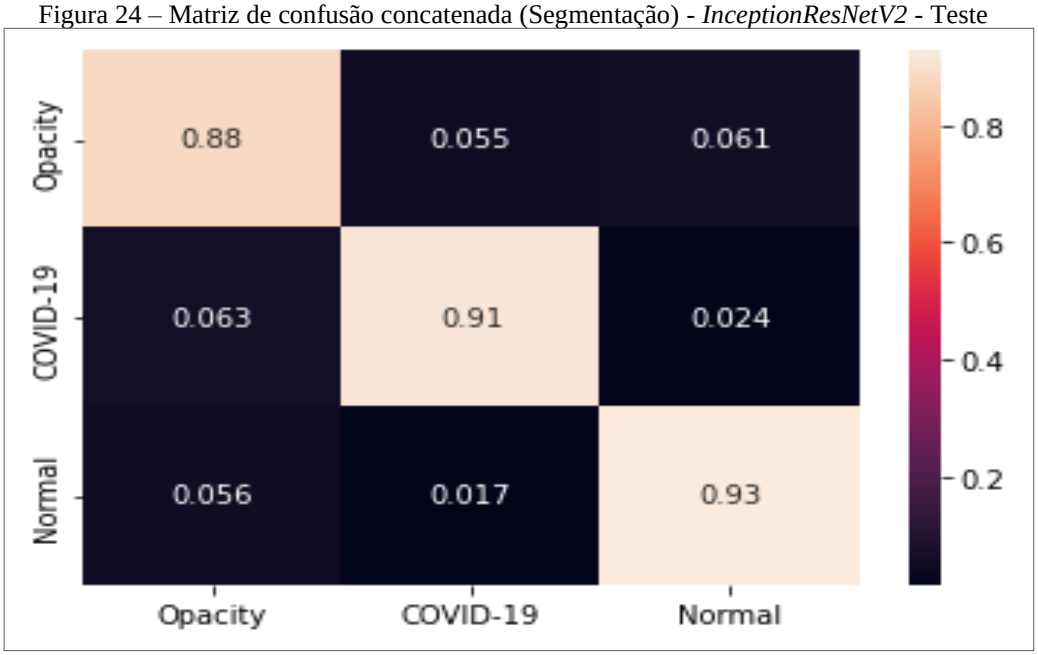
		DenseNet121	InceptionResNetV2	VGG16
Acurácia de teste	média	0,898	0,908	0,899
	desvio padrão	0,009	0,009	0,009
	máximo	0,910	0,923	0,915
Acurácia de teste secundário	média	0,538	0,569	0,482
	desvio padrão	0,042	0,051	0,070
	máximo	0,600	0,635	0,607
Acurácia de treino	média	0,995	0,999	0,977
	desvio padrão	0,009	0,000	0,035
	máximo	0,999	1,000	0,999
F1-Score de teste	média	0,898	0,908	0,899
	desvio padrão	0,009	0,009	0,009
	máximo	0,910	0,923	0,915
F1-Score de teste secundário	média	0,515	0,551	0,448
	desvio padrão	0,046	0,053	0,082
	máximo	0,593	0,627	0,593
F1-Score de treino	média	0,995	0,999	0,977
	desvio padrão	0,009	0,000	0,035
	máximo	0,999	1,000	0,999
Recall de teste	média	0,898	0,908	0,899
	desvio padrão	0,009	0,009	0,009
	máximo	0,910	0,923	0,915
Recall de teste secundário	média	0,538	0,569	0,482
	desvio padrão	0,042	0,051	0,070
	máximo	0,600	0,635	0,607
Recall de treino	média	0,995	0,999	0,977
	desvio padrão	0,009	0,000	0,035
	máximo	0,999	1,000	0,999
Precisão de teste	média	0,899	0,908	0,899
	desvio padrão	0,009	0,009	0,009
	máximo	0,910	0,924	0,915
Precisão de teste secundário	média	0,649	0,670	0,621
	desvio padrão	0,029	0,023	0,053
	máximo	0,714	0,724	0,712
Precisão de treino	média	0,996	0,999	0,977

	desvio padrão	0,009	0,000	0,035
	máximo	0,999	1,000	0,999

Fonte: Própria

A partir dos resultados obtidos com a segmentação, é possível verificar que apesar da acurácia e do *F1-score* de teste terem diminuído de 93,7% para 90,8%, eles obtiveram um aumento de 42,2% para 56,9% na acurácia média para o conjunto de testes secundário. Desse modo, ao realizar a segmentação, a capacidade de generalização do modelo melhorou devido à exclusão de ruídos nas imagens que poderiam gerar enviesamento, indicando que é possível uma menor captura de características específicas das bases e maior captura de características referentes a aparência dos pulmões em relação aos modelos *baseline*. O *F1-score* médio no conjunto de testes secundário saltou de 37,6% para 55,1%, enquanto o seu melhor cenário foi de 57,2% para 62,7%.

Figura 24 – Matriz de confusão concatenada (Segmentação) - *InceptionResNetV2* - Teste



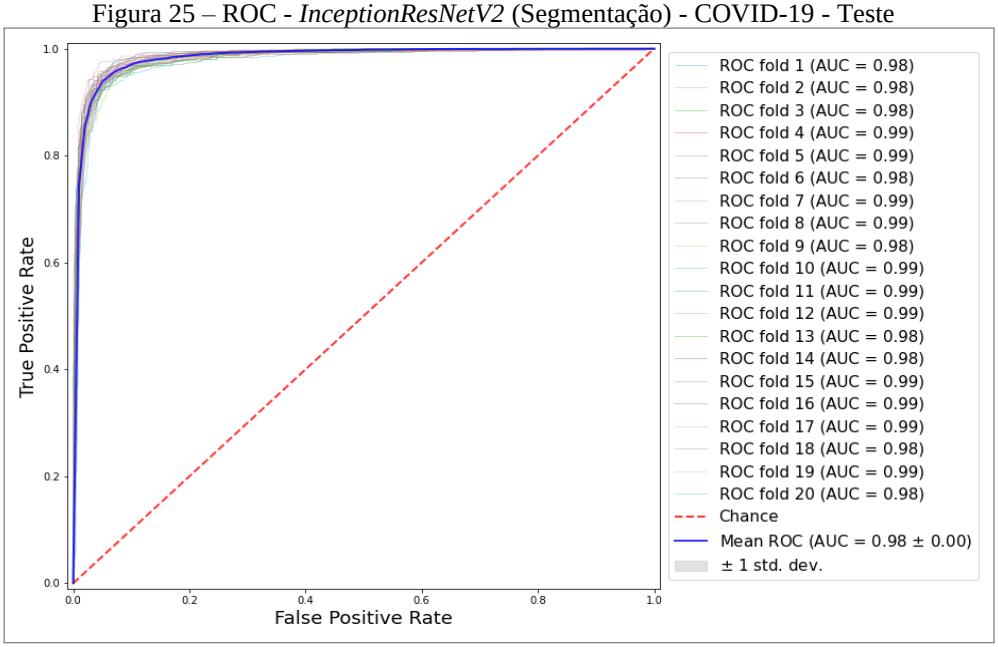
Fonte: Própria

Na figura 24, temos a matriz de confusão concatenada. Percebe-se uma leve alteração no comportamento do modelo, revelando uma classe com maior número de verdadeiros positivos de pulmões normais. Além disso, o modelo não tem como erro predominante a confusão entre a classe normal e pneumonia, uma vez que esse erro é equivalente ao erro de distinção entre as classes COVID-19 e pneumonia.

O comportamento da curva ROC para classe COVID-19 sofre uma leve alteração, refletindo a diminuição de verdadeiros positivos. Na matriz de confusão, verifica-se, também, uma diminuição de 1 ponto percentual na AUC do modelo que utiliza segmentação.

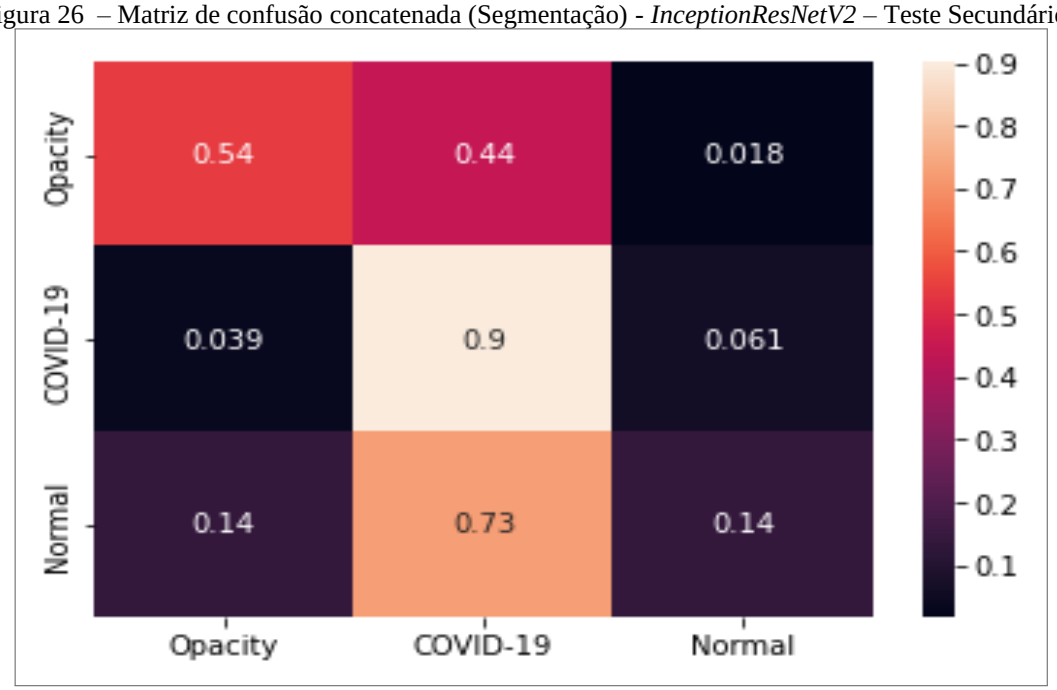
Em relação à matriz de confusão para o conjunto de testes secundários, a maior mudança apresenta um aumento na quantidade de verdadeiros positivos da classe pneumonia. Isso comprova, mais uma vez, a diminuição no enviesamento do modelo.

Figura 25 – ROC - *InceptionResNetV2* (Segmentação) - COVID-19 - Teste



Fonte: Própria

Figura 26 – Matriz de confusão concatenada (Segmentação) - *InceptionResNetV2* – Teste Secundário



Fonte: Própria

4.1.3 *Pipeline* com DCGAN

Tabela 12 – Métricas da *pipeline* com DCGAN para *dataset 1*

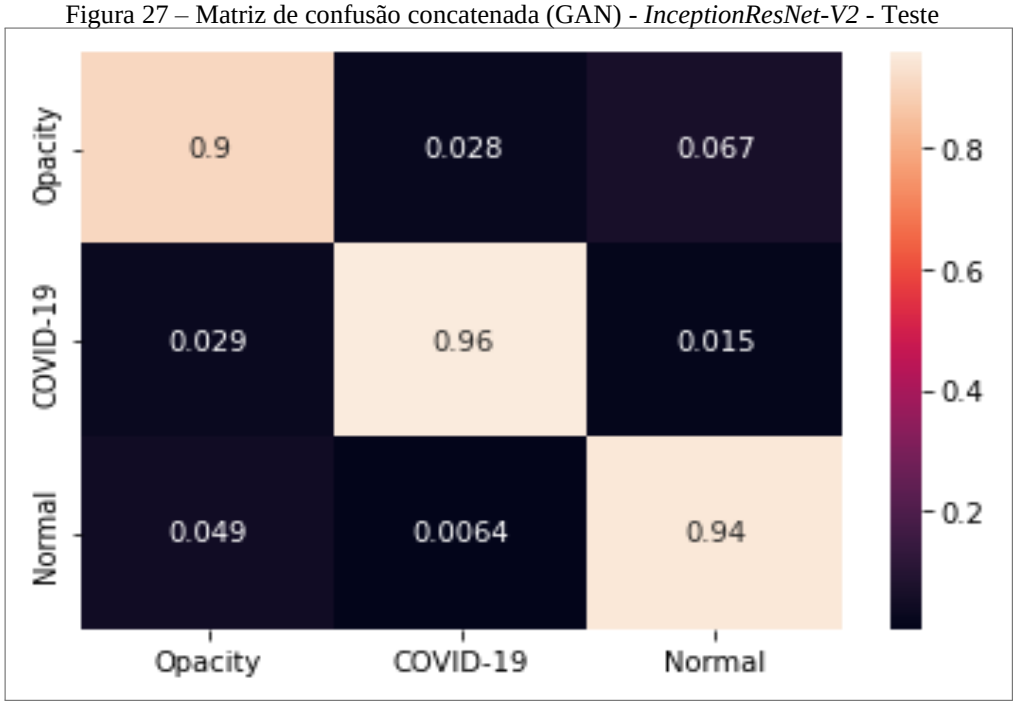
		DenseNet121	InceptionResNetV2	VGG16
Acurácia de teste	média	0,931	0,934	0,932
	desvio padrão	0,005	0,010	0,006
	máximo	0,941	0,944	0,942
Acurácia de teste secundário	média	0,407	0,419	0,408
	desvio padrão	0,046	0,042	0,054
	máximo	0,492	0,525	0,512
Acurácia de treino	média	0,998	0,997	0,995
	desvio padrão	0,001	0,010	0,010
	máximo	1,000	1,000	1,000
F1-Score de teste	média	0,931	0,934	0,932
	desvio padrão	0,005	0,009	0,006
	máximo	0,941	0,944	0,942
F1-Score de teste secundário	média	0,368	0,374	0,358
	desvio padrão	0,066	0,060	0,079
	máximo	0,500	0,513	0,487
F1-Score de treino	média	0,998	0,997	0,995
	desvio padrão	0,001	0,010	0,010
	máximo	1,000	1,000	1,000
Recall de teste	média	0,931	0,934	0,932
	desvio padrão	0,005	0,010	0,006
	máximo	0,941	0,944	0,942
Recall de teste secundário	média	0,407	0,419	0,408
	desvio padrão	0,046	0,042	0,054
	máximo	0,492	0,525	0,512
Recall de treino	média	0,998	0,997	0,995
	desvio padrão	0,001	0,010	0,010
	máximo	1,000	1,000	1,000
Precisão de teste	média	0,931	0,935	0,933
	desvio padrão	0,004	0,009	0,006
	máximo	0,941	0,945	0,942
Precisão de teste secundário	média	0,586	0,621	0,649
	desvio padrão	0,033	0,042	0,027
	máximo	0,641	0,673	0,686
Precisão de treino	média	0,998	0,997	0,995

	desvio padrão	0,001	0,010	0,010
	máximo	1,000	1,000	1,000

Fonte: Própria

A adição de 1500 imagens sintéticas ao conjunto de dados não aparenta ter influenciado de maneira positiva os resultados dos modelos, seja pela qualidade das imagens, visto que GANs são utilizadas para gerar imagens de baixa resolução (MIKOLAJCZYK; GROCHOWSKI, 2018), seja pelo volume de novas imagens que pode não ter sido grande o suficiente, devido às limitações físicas da máquina na realização dos experimentos, para ter um impacto real na diversidade dos dados. Em geral, as métricas se mantiveram semelhantes às *baseline*. No entanto, houve um aumento de desempenho do modelo *InceptionResNet-V2* na base de testes secundária e uma queda de desempenho da VGG16 em relação ao conjunto de testes secundário.

Figura 27 – Matriz de confusão concatenada (GAN) - *InceptionResNet-V2* - Teste



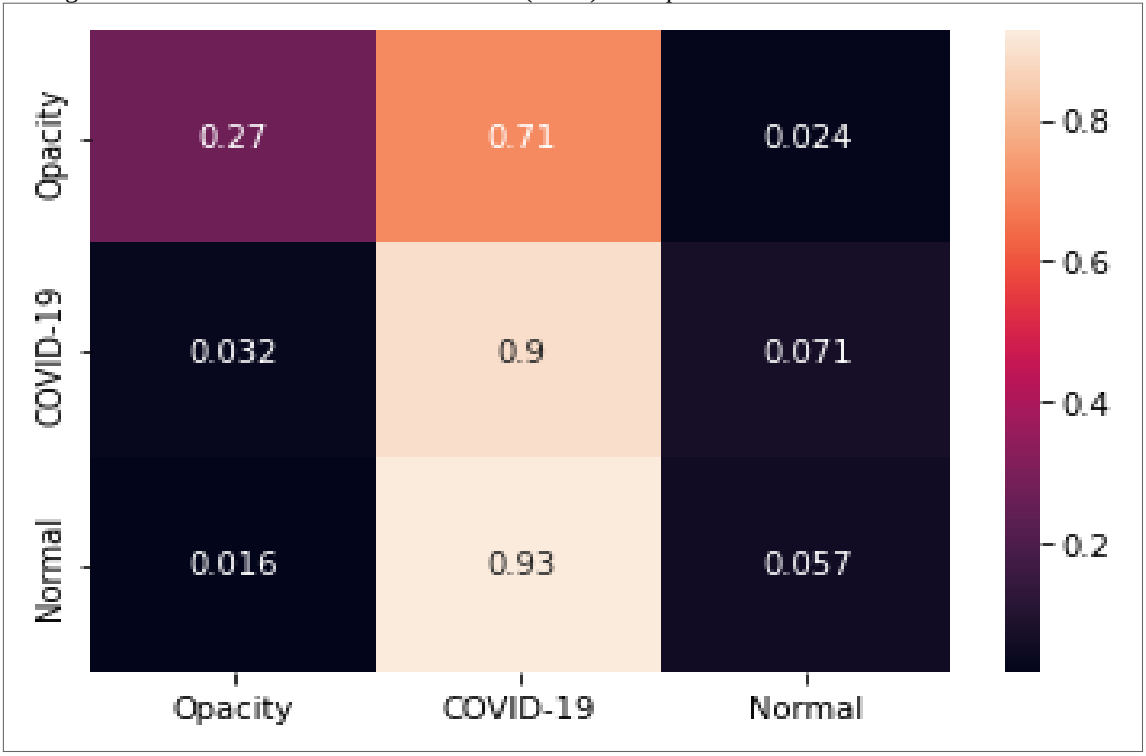
Fonte: Própria

A matriz de confusão para o conjunto de testes segue sendo similar a *baseline*, tendo a classe COVID-19 com o maior número de verdadeiros positivos e o maior erro entre a diferenciação das classes normal e pneumonia.

Pode-se observar, na figura 28, que a adição de dados sintéticos não gerou mudanças drásticas na matriz de confusão para a base de testes secundário ao se comparar com a *baseline*. No entanto, a proporção de verdadeiros positivos para classe normal teve uma queda, enquanto

a de pneumonia teve um aumento. A classe COVID-19 continua concentrando o maior número de predições, mostrando, novamente, um possível enviesamento.

Figura 28 – Matriz de confusão concatenada (GAN) - *InceptionResNet-V2* - Teste Secundário



Fonte: Própria

4.1.4 *Pipeline* completa

Tabela 13 – Métricas da *pipeline* completa para *dataset 1*

		DenseNet121	InceptionResNetV2	VGG16
Acurácia de teste	média	0,972	0,936	0,966
	desvio padrão	0,002	0,003	0,003
	máximo	0,975	0,942	0,970
Acurácia de teste secundário	média	0,250	0,383	0,273
	desvio padrão	0,017	0,033	0,028
	máximo	0,281	0,471	0,335
Acurácia de treino	média	0,998	0,999	0,998
	desvio padrão	0,001	0,001	0,004
	máximo	1,000	1,000	1,000
F1-Score de teste	média	0,972	0,936	0,966

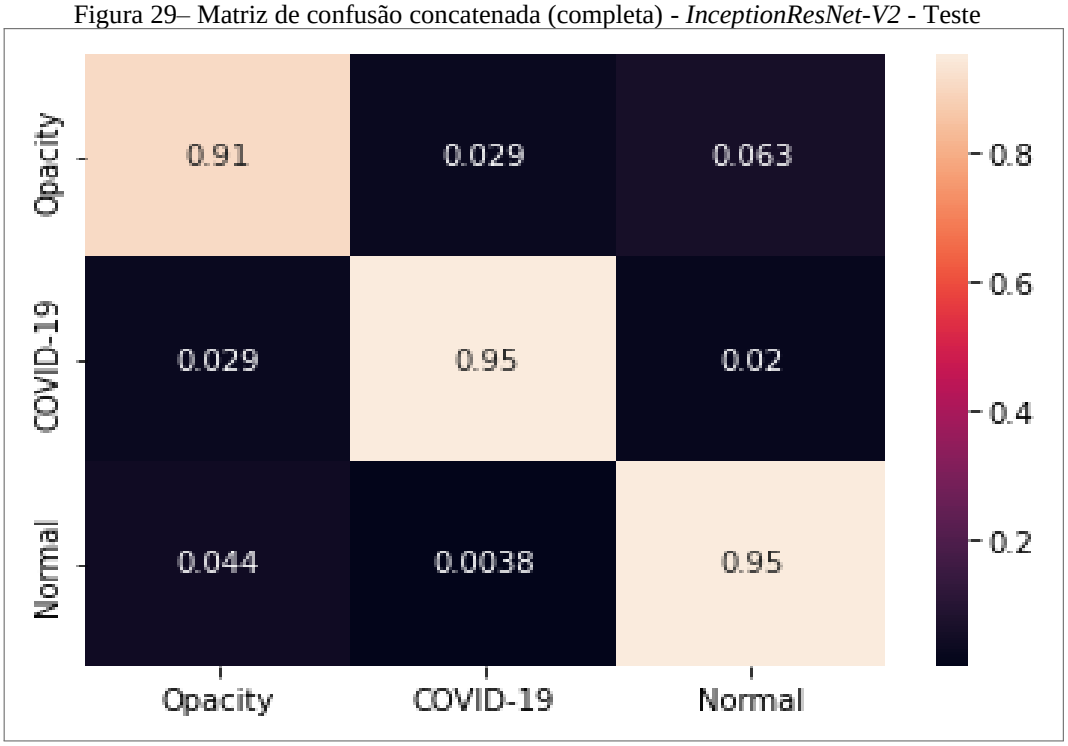
F1-Score de teste secundário	desvio padrão	0,002	0,003	0,003
	máximo	0,975	0,942	0,970
	média	0,197	0,325	0,230
	desvio padrão	0,023	0,053	0,043
	máximo	0,229	0,451	0,312
F1-Score de treino	média	0,998	0,999	0,998
	desvio padrão	0,001	0,001	0,004
	máximo	1,000	1,000	1,000
Recall de teste	média	0,972	0,936	0,966
	desvio padrão	0,002	0,003	0,003
	máximo	0,975	0,942	0,970
Recall de teste secundário	média	0,250	0,383	0,273
	desvio padrão	0,017	0,033	0,028
	máximo	0,281	0,471	0,335
Recall de treino	média	0,998	0,999	0,998
	desvio padrão	0,001	0,001	0,004
	máximo	1,000	1,000	1,000
Precisão de teste	média	0,972	0,937	0,966
	desvio padrão	0,002	0,003	0,003
	máximo	0,976	0,942	0,970
Precisão de teste secundário	média	0,566	0,622	0,583
	desvio padrão	0,021	0,022	0,027
	máximo	0,617	0,690	0,652
Precisão de treino	média	0,998	0,999	0,998
	desvio padrão	0,001	0,001	0,004
	máximo	1,000	1,000	1,000

Fonte: Própria

Com a *pipeline* completa, isto é, realizando a criação de dados sintéticos e a segmentação de imagens, observa-se que há uma melhora significativa nas métricas das redes *DenseNet121* e VGG-16, enquanto a *InceptionResNet-V2* permanece com valores similares aos *baseline*. Todavia, em relação ao desempenho na base de testes secundária, há uma piora para

todos os modelos, sendo o *InceptionResNet-V2* o menos afetado. É possível que, ao combinar ambas as técnicas, tenha sido criado o efeito contrário do que se buscava e o enviesamento dos modelos aumentou, justificando o aumento de acurácia no teste e sua diminuição no teste secundário.

Figura 29– Matriz de confusão concatenada (completa) - *InceptionResNet-V2* - Teste

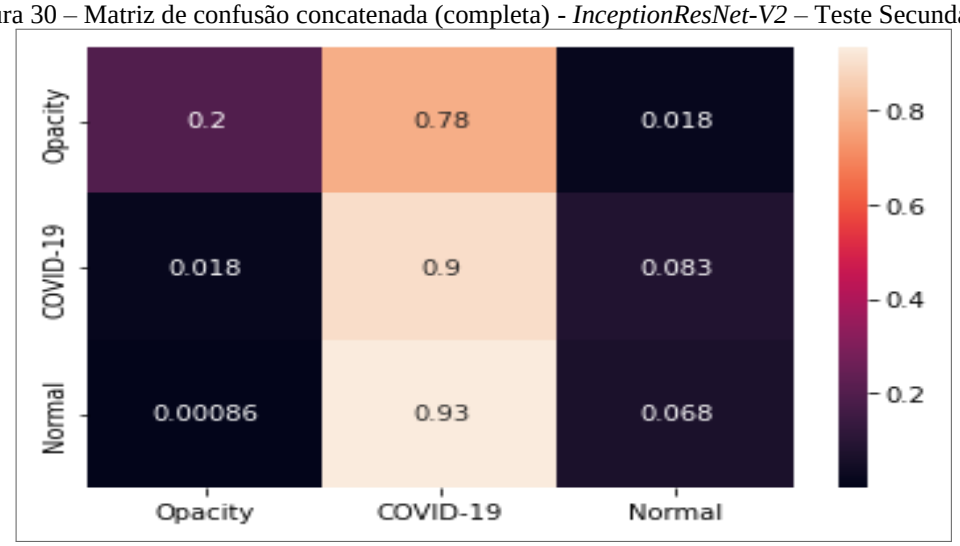


Fonte: Própria

A matriz de confusão para o conjunto de testes tem o comportamento similar a *baseline*, uma vez que a classe COVID-19 tem o maior número de verdadeiros positivos, juntamente com a classe normal. O maior erro continuou apresentando-se na diferenciação entre as classes normal e pneumonia, tendo sua diferença mais pronunciada em relação a *baseline*, na qual a proporção de verdadeiros positivos aumentou em todas as classes.

A figura 30 aponta que a adição de dados sintéticos e segmentação apenas diminuiu a quantidade de verdadeiros positivos na base de testes secundária para todas as classes ao se comparar com a *baseline*. Em relação à *pipeline* anterior, uma única mudança positiva é encontrada no número de verdadeiros positivos para classe normal, que sofre um leve aumento. Assim como nos outros modelos a classe COVID-19, continua concentrando o maior número de predições.

Figura 30 – Matriz de confusão concatenada (completa) - *InceptionResNet-V2* – Teste Secundário



Fonte: Própria

4.2 RESULTADOS *DATASET* 2

4.2.1 *Pipeline baseline*

Tabela 14 – Métricas da *pipeline* baseline para *dataset 2*

		DenseNet121	VGG16
Acurácia de teste	média	0,968	0,964
	desvio padrão	0,005	0,006
	máximo	0,974	0,971
Acurácia de teste secundário	média	0,259	0,254
	desvio padrão	0,030	0,038
	máximo	0,316	0,343
Acurácia de treino	média	0,996	0,993
	desvio padrão	0,007	0,011
	máximo	1,000	1,000
F1-Score de teste	média	0,968	0,964
	desvio padrão	0,005	0,006
	máximo	0,974	0,971
F1-Score de teste secundário	média	0,202	0,206
	desvio padrão	0,036	0,049
	máximo	0,263	0,298
F1-Score de treino	média	0,996	0,993
	desvio padrão	0,007	0,011

	máximo	1,000	1,000
Recall de teste	média	0,968	0,964
	desvio padrão	0,005	0,006
	máximo	0,974	0,971
Recall de teste secundário	média	0,259	0,254
	desvio padrão	0,030	0,038
	máximo	0,316	0,343
Recall de treino	média	0,996	0,993
	desvio padrão	0,007	0,011
	máximo	1,000	1,000
Precisão de teste	média	0,968	0,964
	desvio padrão	0,005	0,006
	máximo	0,975	0,972
Precisão de teste secundário	média	0,525	0,585
	desvio padrão	0,090	0,054
	máximo	0,632	0,654
Precisão de treino	média	0,996	0,993
	desvio padrão	0,007	0,011
	máximo	1,000	1,000

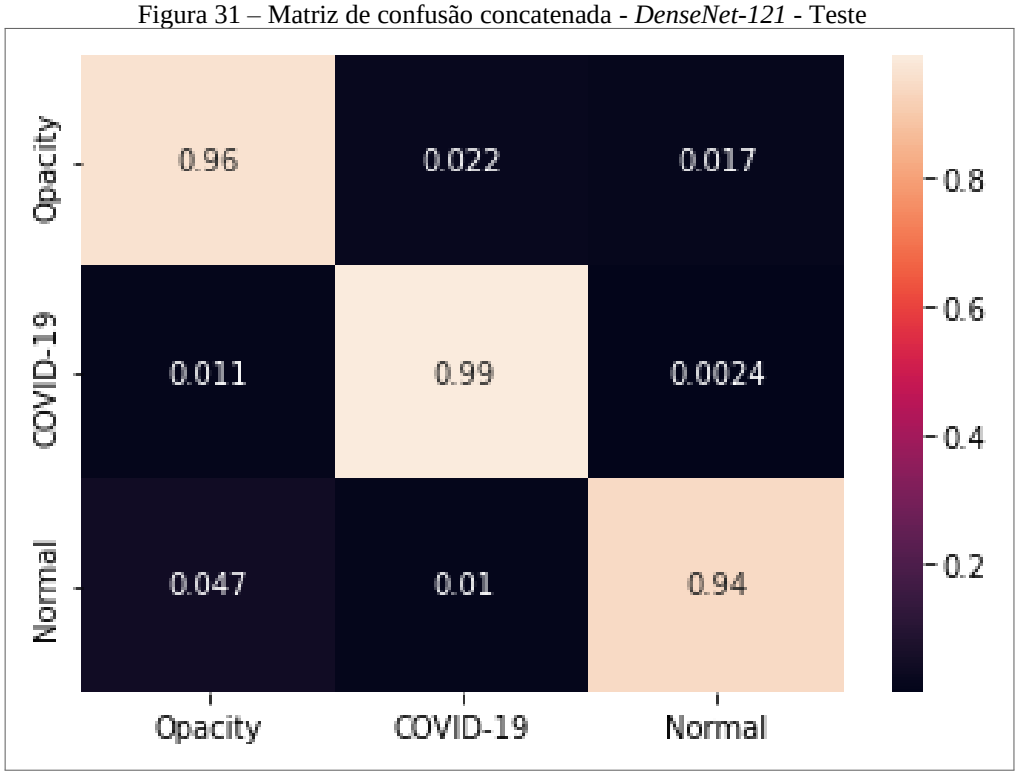
Fonte: Própria

Ao utilizar o *dataset* 2, verifica-se que a acurácia e o *F1-score* para este conjunto de dados crescem em alguns pontos percentuais comparado aos modelos *baselines* do *dataset* 1. No entanto, o mesmo não acontece com essas métricas para o conjunto de testes secundários, isto é, eles só apresentam alta performance em imagens com características bastante semelhantes as que estão na base de treino.

Isso pode ter acontecido devido à base OCT ser proveniente de um hospital infantil, o que faz com que existam mais diferenças entre as classes COVID-19, normal e pneumonia, os quais não se restringem apenas a alterações na área pulmonar, na medida em que o tamanho e formato dos pulmões pode estar influenciando na classificação e tornando “mais fácil” a tarefa do modelo de diferenciar as classes. Dessa forma, há um maior enviesamento dos modelos, e, portanto, diminuição da capacidade de generalização.

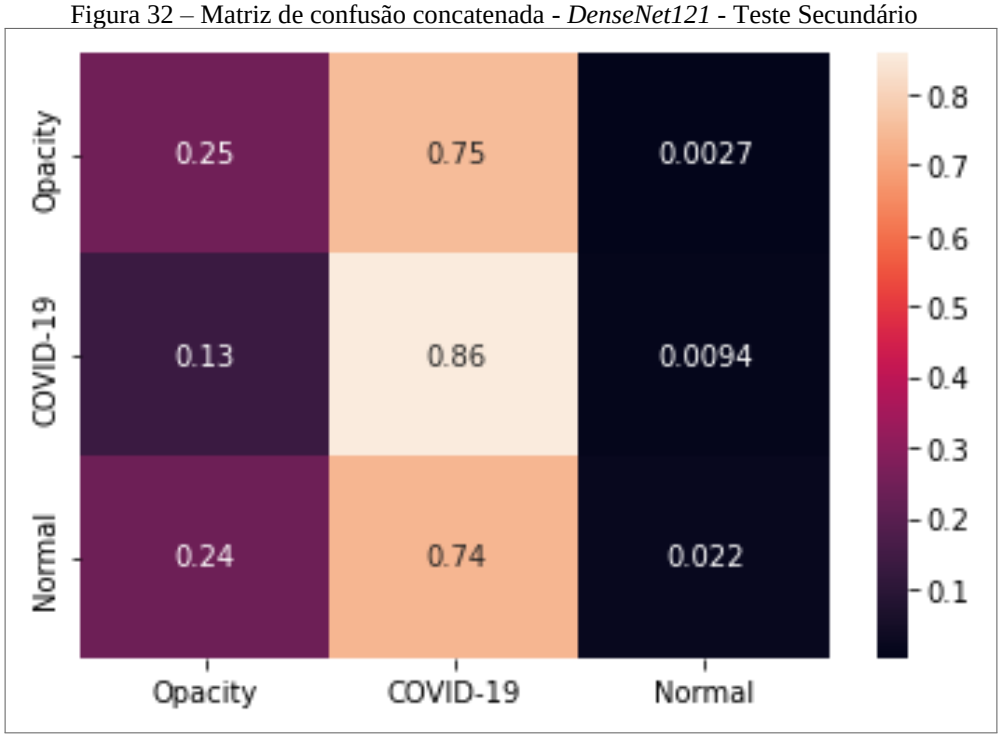
A matriz de confusão concatenada dos experimentos, apresentada na figura 31, corrobora para a hipótese levantada sobre o viés das bases utilizadas, visto que o modelo consegue separar quase que de maneira perfeita a classe COVID-19 das demais.

Figura 31 – Matriz de confusão concatenada - *DenseNet-121* - Teste



Fonte: Própria

Figura 32 – Matriz de confusão concatenada - *DenseNet121* - Teste Secundário



Fonte: Própria

A matriz de confusão concatenada dos resultados da base de testes secundária, na figura 32, mostra a classe COVID-19 com a maior concentração de predições, indicando mais uma

vez um possível enviesamento. A maior diferença entre ela e as matrizes geradas pelos modelos treinados no *dataset* 1 é que também há certa concentração de predições para a classe pneumonia, fazendo com que quase não sejam detectadas imagens pertencentes a classe normal.

4.2.2 *Pipeline* com segmentação

Tabela 15 – Métricas da *pipeline* segmentado para *dataset 2*

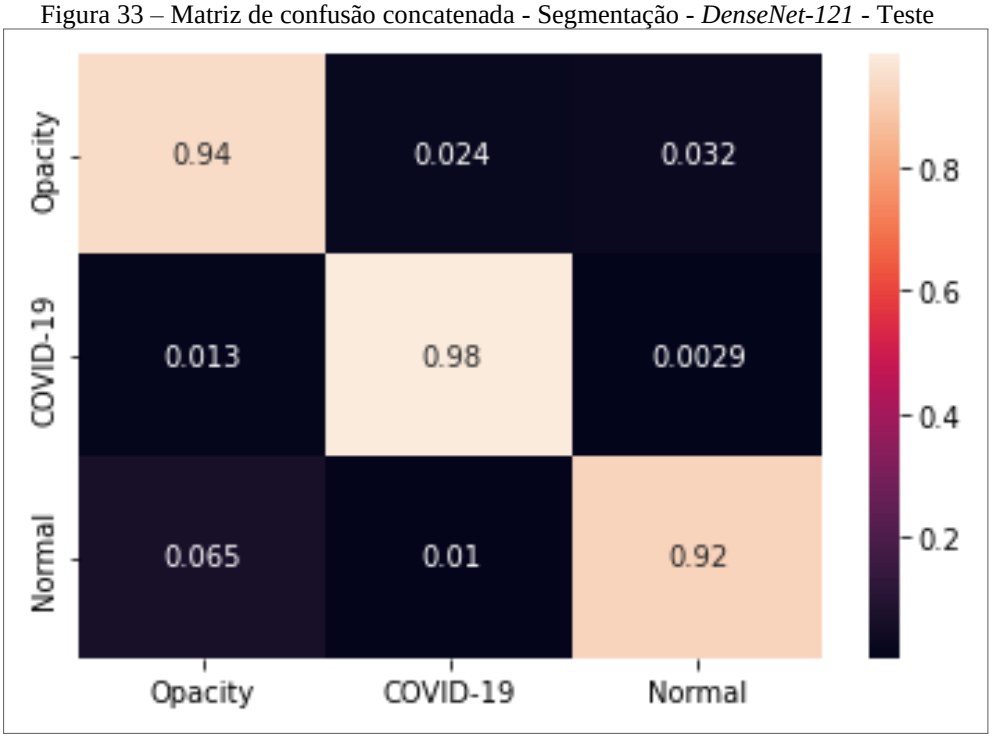
		DenseNet121
Acurácia de teste	média	0,957
	desvio padrão	0,006
	máximo	0,968
Acurácia de teste secundário	média	0,221
	desvio padrão	0,022
	máximo	0,266
Acurácia de treino	média	0,995
	desvio padrão	0,009
	máximo	1,000
F1-Score de teste	média	0,957
	desvio padrão	0,006
	máximo	0,968
F1-Score de teste secundário	média	0,151
	desvio padrão	0,038
	máximo	0,225
F1-Score de treino	média	0,995
	desvio padrão	0,009
	máximo	1,000
Recall de teste	média	0,957
	desvio padrão	0,006
	máximo	0,968
Recall de teste secundário	média	0,221
	desvio padrão	0,022
	máximo	0,266
Recall de treino	média	0,995

	desvio padrão	0,009
	máximo	1,000
Precisão de teste	média	0,957
	desvio padrão	0,006
	máximo	0,968
Precisão de teste secundário	média	0,629
	desvio padrão	0,083
	máximo	0,734
Precisão de treino	média	0,995
	desvio padrão	0,009
	máximo	1,000

Fonte**:** Própria

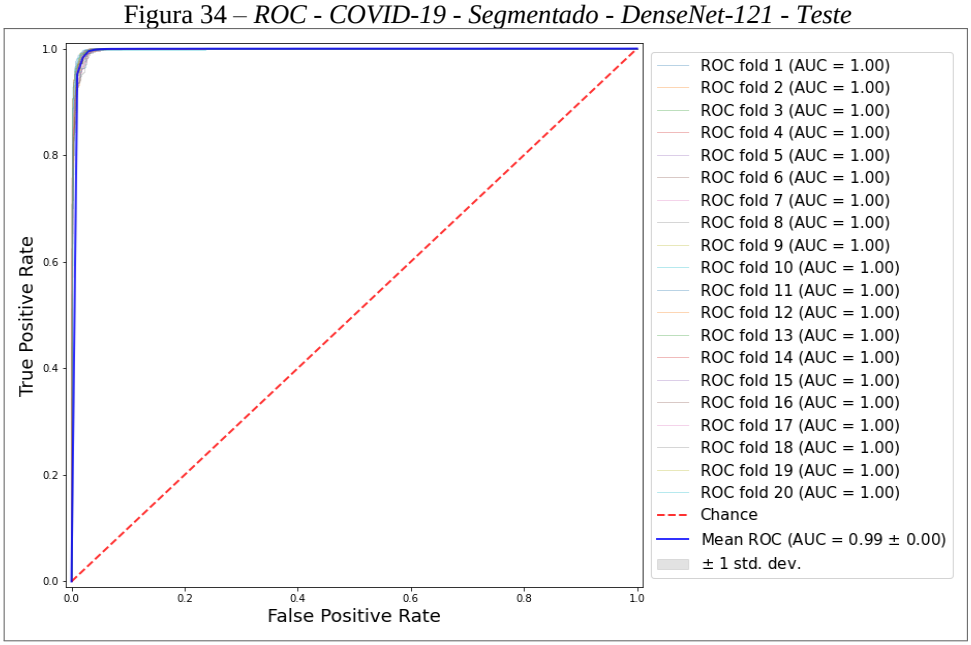
Ao aplicar a segmentação no conjunto de dados 2, observa-se que as acurácias e *F1-scores* diminuem como esperado, no entanto, um comportamento novo acontece ao verificar as métricas do conjunto de testes secundário. Diferente do que ocorreu na mesma aplicação no conjunto de dados 1, houve também uma piora das métricas.

Figura 33 – Matriz de confusão concatenada - Segmentação - *DenseNet-121* - Teste



Fonte: Própria

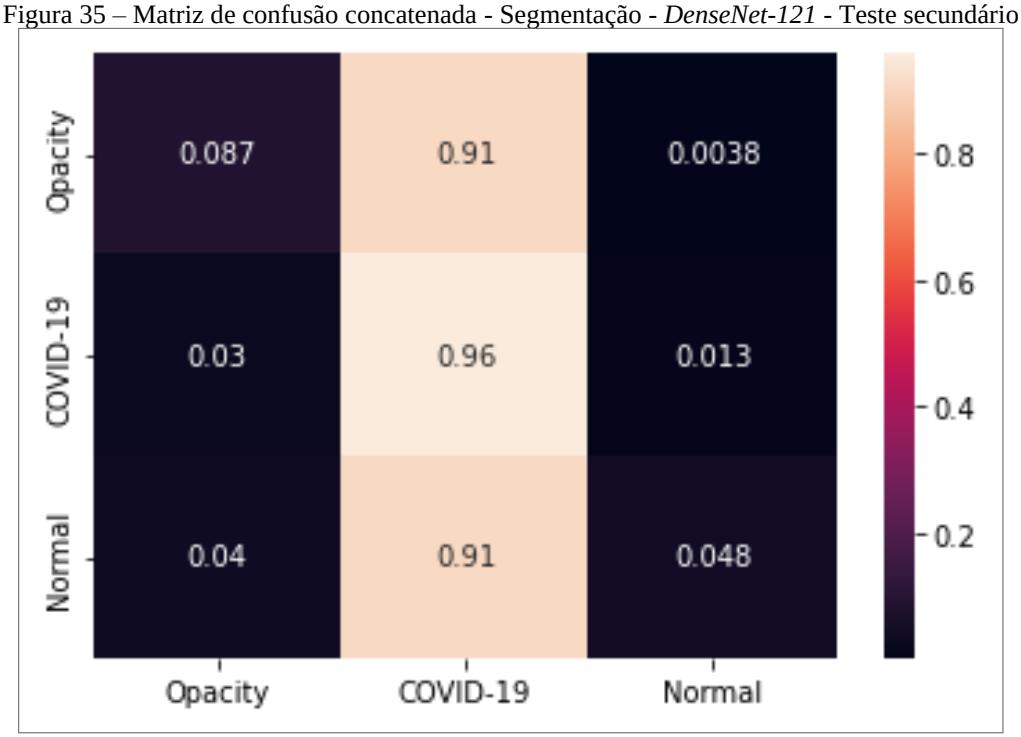
Figura 34 – *ROC - COVID-19 - Segmentado - DenseNet-121 - Teste*



Fonte: Própria

A matriz de confusão concatenada dos experimentos, figura 34, e a curva ROC da classe COVID-19, figura 35, mostram uma alta distinção da classe COVID-19 em relação às demais, assim como na ocorre na pipeline *baseline*.

Figura 35 – Matriz de confusão concatenada - Segmentação - *DenseNet-121* - Teste secundário



Fonte: Própria

A matriz de confusão concatenada dos experimentos na base de testes secundária, mostra um desempenho ainda pior do que o visto na *baseline*, quase todas as predições foram

concentradas na classe COVID-19, apresentando que para este conjunto de dados a segmentação não funcionou como fator mitigante do viés entre as bases.

Com estes resultados é possível visualizar a importância da diversidade de dados ao realizar o treinamento de modelos de *machine learning*, dado que quando utilizamos um conjunto pouco diverso, como o OCT, que apresenta apenas dados de um hospital voltado ao público infantil, tem-se resultados enganosos e um modelo com acurácia aumentada artificialmente (MAIOR et al., 2021).

Também observa-se a importância de realizar repetições no lugar de apresentar apenas o melhor resultado do modelo, visto que em alguns casos a diferença entre a performance média do modelo e a performance máxima vai até 14 pontos percentuais para o conjunto de testes secundários e até 2 pontos percentuais no conjunto de testes padrão. Dessa maneira, ao apresentar apenas os melhores modelos, é possível que esteja apontando uma métrica de avaliação que não corresponde com a realidade dos modelos.

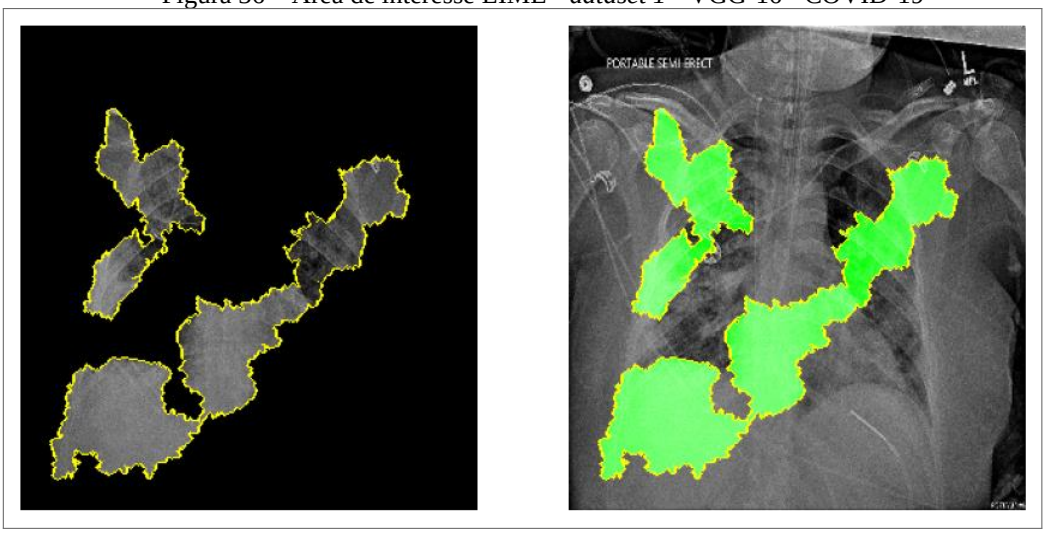
4.3 APLICAÇÃO DO LIME

Foi utilizada a biblioteca LIME (Local *Interpretable Model-agnostic Explanations*) (RIBEIRO; SINGH; GUESTRIN, 2016) para interpretar o comportamento dos melhores modelos das *pipelines baseline* e segmentadas e entender que características estão sendo procuradas por cada uma delas.

Na figura 36, há uma interpretação de área de interesse do modelo VGG-16 treinado com o *dataset* 1. Na imagem recortada, destaca-se áreas da imagem que influenciaram o modelo a escolher corretamente COVID-19 como a classe a qual esta imagem pertence, sendo viável observar as mesmas áreas em verde na imagem completa do raio-X.

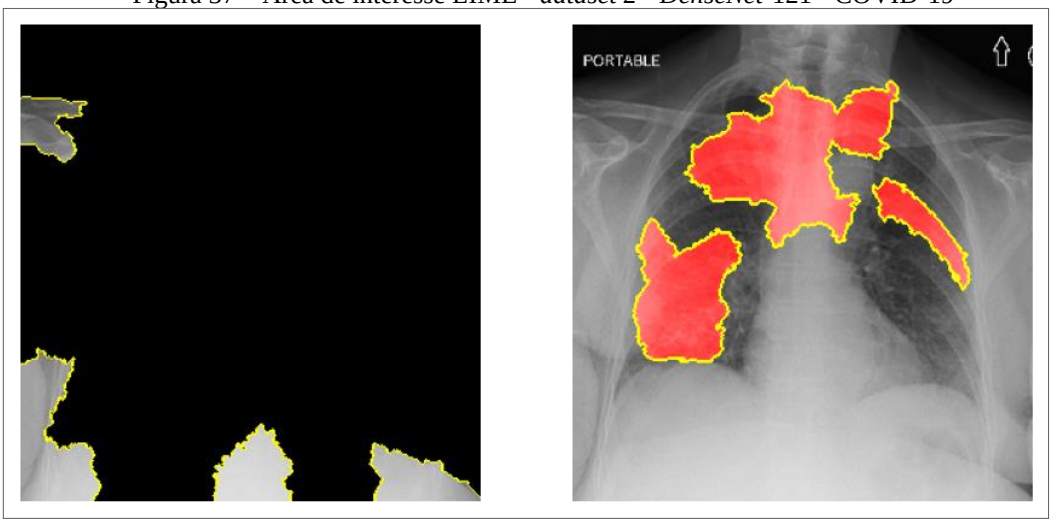
Para o modelo *DenseNet-121* treinado no *dataset* 2, na figura 37, observa-se que as áreas em destaque na imagem recortada não são pertencentes ao pulmão, no entanto elas influenciaram o modelo positivamente na escolha da classe correta para imagem. Em vermelho, estão as áreas da imagem que diminuem a probabilidade de ela ser classificada como COVID-19.

Figura 36 – Área de interesse LIME - *dataset 1* - VGG-16 - COVID-19



Fonte: Própria

Figura 37 – Área de interesse LIME - *dataset 2* - *DenseNet*-121 - COVID-19



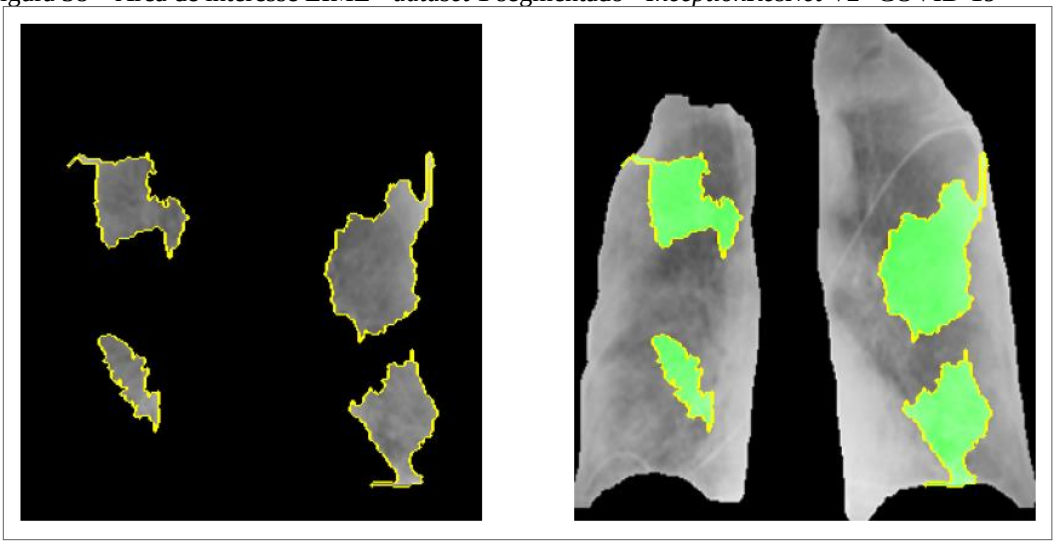
Fonte: Própria

As áreas de interesse do modelo *InceptionResNet-V2* treinado com o *dataset* 1, são visualizadas na figura 38. Em ambas as imagens, observar-se em destaque as áreas do raio-X que influenciaram o modelo a escolher corretamente COVID-19 como a classe a qual ele pertence. É importante notar que, no modelo sem a segmentação, foram destacadas áreas não pertencentes ao pulmão do paciente, enquanto no modelo com segmentação há uma limitação na área analisada da imagem, o que força o modelo a procurar extrair características apenas dos pulmões dos pacientes.

O mesmo comportamento pode ser observado ao se analisar a figura 39, e compará-la com a 37. No *dataset* 2, a segmentação também foi importante para limitar a área de extração

de características e tornar o modelo mais confiável, mesmo que isso implique em uma queda de performance.

Figura 38 – Área de interesse LIME - *dataset* 1 segmentado - *InceptionResNet-V2* -COVID-19



Fonte: Própria

Figura 39 – Área de interesse LIME - *dataset* 2 segmentado - *DenseNet*-121 - COVID-19

Fonte: Própria

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONSIDERAÇÕES FINAIS

Neste estudo foram criados 4 *pipelines* distintos para fins de comparação de combinações de 3 técnicas de *deep learning* na tarefa de classificação multiclasse de radiografias de pacientes com COVID-19, pneumonia não proveniente da COVID-19 e pulmões saudáveis. Além do objetivo de encontrar um *pipeline* ótimo para a classificação, o presente trabalho oferece uma discussão sobre as possíveis vantagens de cada método utilizado e procura expor uma lacuna na literatura em relação a análise da capacidade de generalização dos modelos criados e propostos até então.

O *pipeline baseline* apresenta aumento de dados e treina uma rede CNN, a qual pode ser a VGG-16, *DenseNet* ou *InceptionResNet-V2*, por meio de *transfer learning* com um conjunto de treino aumentado, testando-a com 2 conjuntos de dados. As bases de dados de testes são: um conjunto de testes separado do treino antes da etapa de treinamento do modelo; um conjunto intitulado de teste secundária, o qual contém bases completamente distintas daquelas que foram utilizadas para treinar o modelo, de forma a simular uma aplicação real, em que os dados de *input* podem ter características bastante diferentes em comparação aos dados de treinamento. Dessa maneira, buscou-se avaliar tanto as métricas dos modelos e sua relação com a base de testes convencional, quanto a base de testes secundária, a qual foi utilizada para medir a capacidade de generalização do modelo.

Os *pipelines* restantes utilizam os mesmos procedimentos do *baseline*, também operando *transfer learning* para treinar as CNNs escolhidas, mas aplicam novas técnicas. O segundo *pipeline* utiliza segmentação das imagens por meio da rede *U-Net*, enquanto o terceiro faz a utilização de dados sintéticos criados por uma DCGAN. Por fim, o último *pipeline* combina as 3 técnicas mencionadas para criação de um *pipeline* completo.

Foram realizados 15 experimentos, com 20 repetições cada, seguindo as configurações distintas mencionadas. Os experimentos foram aplicados em 2 conjuntos de dados, os *datasets* 1 e 2, os quais foram obtidos a partir de 8 bases de dados abertos. Dessa maneira, foi feita uma análise sobre o impacto de diferentes técnicas aplicadas aos modelos de redes neurais convolucionais, com a finalidade de melhorar seu desempenho e diminuir seu viés. 12 experimentos foram realizados utilizando o *dataset* 1, enquanto 3 foram realizados com o *dataset* 2 para treinar os modelos.

A partir das métricas calculadas utilizando os conjuntos de testes secundários foi feita, também, uma análise sobre a capacidade de generalização dos modelos. Assim, conclui-se que ainda são necessárias melhorias para que os modelos se tornem capazes de atuar em hospitais como forma de apoio aos médicos. Também foi concluído que a melhor forma, dentre as testadas no presente estudo, de diminuir o viés de um modelo e melhorar sua generalização é pela segmentação de imagens.

Para possibilitar a aplicação dos modelos existentes em hospitais, a fim de auxiliar profissionais da saúde, é necessário que se faça mais análises sobre a capacidade de generalização. Mesmo obtendo acurácias e *F1-Scores* acima de 90%, os estudos até então publicados na literatura não exploram a possibilidade de estarem limitados para dados de bases diferentes. É imprescindível mais diversidade de bases de dados abertas, como também a ampliação de estudos sobre aplicações de técnicas para diminuir possíveis enviesamentos, utilizando redes neurais para detecção da COVID-19 aplicadas em ambientes hospitalares.

Desse modo, o presente estudo não foi capaz de sanar este problema, dado que o melhor *F1-Score* obtido para os testes secundários foi de 62,7%, enquanto o *F1-Score* médio na melhor *pipeline* encontrada foi de 55,1%. No entanto, esta pesquisa pode servir como base para futuros estudos que procurem explorar outras maneiras de solucionar o problema da generalização.

No que se tange ao uso de GANs para criar dados sintéticos, não foram obtidas respostas claras sobre o real impacto da utilização desses dados criados pelas redes generativas, na diversificação dos conjuntos de dados selecionados, ou sobre a aplicabilidade dessa técnica em conjunto com a segmentação de dados. Pode-se concluir, no entanto, que a adição de apenas 500 novas imagens para cada classe não gerou mudanças significativas no desempenho do modelo para que fosse justificável o esforço empregado em realizar esta tarefa. Em outros estudos (LOEY; SMARANDACHE; KHALIFA, 2020; CAVALCANTI; BERTON, 2021), GANs se mostraram ferramentas importantes para o aumento do desempenho dos modelos, portanto, experimentos futuros com estas redes não devem ser descartados.

Também é possível concluir que, considerando todos os pipelines executados, o melhor resultado foi obtido com a segmentação na base de dados 1, pois apesar de não ter alcançado a melhor acurácia ou *F1-Score,* dentre os experimentos, sua capacidade de generalização é a maior comparado aos modelos executados nas *pipelines* deste trabalho, apresentando um *F1-Score* de 55,1% para o conjunto de testes secundário, devido à eliminação de características feita pela segmentação das imagens.

No entanto, sua acurácia e *F1-Score* estão dentro da média dos modelos vistos na literatura, como é possível observar ao analisar a tabela 15 abaixo. Tendo um desempenho

superior aos trabalhos de Loey et al. (2020), Nikolaou et al. (2021), Teixeira et al. (2021) e próximo ao desempenho obtido por Maior et al. (2021), com *F1-Score* de 90,8% para o conjunto de testes a *InceptionResNet-V2*. Juntamente com a segmentação de imagens realizada pela *U-Net*, corresponde a uma *pipeline* ótima proposta por este estudo. Esta escolha também é embasada pela análise de áreas de interesse dos modelos que utilizam o LIME, o qual mostra uma escolha de características mais coerente pelo modelo segmentado em comparação ao modelo *baseline*.

Tabela 16 – Comparativo do modelo proposto com publicações relevantes

Trabalho	F1-Score	Acurácia
(CHAKRABORTY; MURALI; MITRA, 2022)	93,00%	96,43%
(BHATTACHARYYA et al., 2022)	95,00%	96,60%
(TEIXEIRA et al., 2021)	88,00%	-
(NIKOLAOU et al., 2021)	90,00%	-
(MAIOR et al., 2021)	-	91,21%
(NEFOUSSI; AMAMRA; AMAROUCHE, 2021)	94,00%	94,00%
(UMER et al., 2021)	95,51%	89,85%
(RAJKUMAR et al., 2020)	-	96,00%
(KHAN; SHAH; BHAT, 2020)	95,60%	95,00%
(WANG; LIN; WONG, 2020)	-	93,30%
(ABBAS; ABDELSAMEA; GABER, 2020)	-	93.1%
(UCAR; KORKMAZ, 2020)	98,25%	98,26%
(APOSTOLOPOULOS; MPESIANA, 2020)	-	94,72%
(LOEY; SMARANDACHE; KHALIFA, 2020)	85,19%	85,19%
Modelo proposto (desempenho médio)	90,80%	90,80%
Modelo proposto (melhor desempenho)	92,30%	92,30%

Fonte: Própria

Ao avaliar os resultados deste trabalho, também foi possível observar a importância de realizar repetições para obtenção das métricas de avaliação dos modelos, a fim de diminuir o aumento artificial de métricas dos modelos propostos, dado que pode haver diferenças de até 1,5 pontos percentuais entre a melhor performance de um modelo e sua performance média na base de testes. Desse modo, considerando tudo que foi exposto, vale ressaltar que o trabalho conseguiu responder as perguntas de pesquisa elencadas e, portanto, cumpriu com seus objetivos traçados.

5.2 LIMITAÇÕES DO ESTUDO

Devido às limitações físicas de GPU, CPU e à memória da máquina utilizada, não foi possível abranger mais configurações de conjuntos de dados, assim como foi inviável a execução dos modelos com mais de 500 imagens sintéticas adicionadas por questões de tempo de execução. A diversidade dos experimentos também foi limitada devido à máquina disponível para realizar o estudo, visto que poderiam ter sido comparadas mais de uma técnica de segmentação e geração de imagens por meio de GANs.

Além disso, também vale pontuar a escassez de dados abertos disponíveis. Apesar do crescimento acerca da disponibilidade de dados abertos referentes a COVID-19 desde 2020, só existem apenas 2 bases de tamanho considerável e de fácil acesso para imagens de raio-X de pacientes com pneumonia, que foram trabalhadas neste estudo. Uma maior diversidade de dados neste quesito poderia melhorar consideravelmente o desempenho dos modelos para bases externas.

5.3 TRABALHOS FUTUROS

Com base nos resultados descritos, sugere-se o desenvolvimento de pesquisas que utilizem outras configurações de dados, mesclando de maneiras diferentes as 8 bases de dados e criando bases novas de testes secundários. Também recomenda-se o uso de todas bases de dados coletadas para realizar o treinamento dos modelos, bem como a busca de novas fontes para realizar testes secundários, mesclando os conjuntos de dados de OCT e RSNA de maneira a gerar maior diversificação para as bases de pneumonia e pulmão normal. Dados de hospitais locais também poderiam ser utilizados como conjunto de testes secundário.

Por fim, recomenda-se, também, a realização de experimentos que explorem outras GANs, gerando uma quantidade maior de dados sintéticos, e outras redes de segmentação, de maneira que seja possível comparar sua performance com a dos experimentos deste estudo.

REFERÊNCIAS

ABBAS, A.; ABDELSAMEA, M. M.; GABER, M. M. **Classification of COVID-19 in chest x-ray images using DeTraC deep convolutional neural network**. *Applied Intelligence*, Springer Science and Business Media LLC, v. 51, n. 2, p. 854–864, set. 2020. Disponível em: <https://doi.org/10.1007/s10489-020-01829-7>.

AI, T.; YANG, Z.; HOU, H.; ZHAN, C.; CHEN, C.; LV, W.; TAO, Q.; SUN, Z.; XIA, L. **Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in china**: A report of 1014 cases. *Radiology*, Radiological Society of North America (RSNA), v. 296, n. 2, p. E32–E40, ago. 2020. Disponível em: <https://doi.org/10.1148/radiol.2020200642>.

ANAYA-ISAZA, A.; MERA-JIMéNEZ, L.; ZEQUERA-DIAZ, M. **An overview of deep learning in medical imaging**. *Informatics in Medicine Unlocked*, v. 26, p. 100723, 2021. ISSN 2352-9148. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2352914821002033>.

APOSTOLOPOULOS, I. D.; MPESIANA, T. A. **Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks**. *Physical and Engineering Sciences in Medicine*, Springer Science and Business Media LLC, v. 43, n. 2, p. 635–640, abr. 2020. Disponível em: <https://doi.org/10.1007/s13246-020-00865-4>.

Front matter. In: BAJAJ, V.; SINHA, G. (Ed.). ***Artificial Intelligence-Based Brain Computer Interface***. Academic Press, 2022. p. i–ii. ISBN 978-0-323-91197-9. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780323911979099926>.

BERNHEIM, A.; MEI, X.; HUANG, M.; YANG, Y.; FAYAD, Z. A.; ZHANG, N.; DIAO, K.; LIN, B.; ZHU, X.; LI, K.; LI, S.; SHAN, H.; JACOBI, A.; CHUNG, M. **Chest CT findings in coronavirus disease-19 (COVID-19)**: Relationship to duration of infection. *Radiology*, Radiological Society of North America (RSNA), v. 295, n. 3, p. 200463, jun. 2020. Disponível em: <https://doi.org/10.1148/radiol.2020200463>.

BHATTACHARYYA, A.; BHAIK, D.; KUMAR, S.; THAKUR, P.; SHARMA, R.; PACHORI, R. B. **A deep learning based approach for automatic detection of COVID-19 cases using chest x-ray images**. *Biomedical Signal Processing and Control*, Elsevier BV, v. 71, p. 103182, jan. 2022. Disponível em: <https: //doi.org/10.1016/j.bspc.2021.103182>.

BULL-OTTERSON, L.; BACA, S.; SAYDAH, S.; BOEHMER, T. K.; ADJEI, S.; GRAY, S.; HARRIS, A. M. **Post–COVID conditions among adult COVID-19 survivors aged 18–64 and ≥65 years — united states, march 2020–november 2021**. MMWR. Morbidity

and Mortality Weekly Report, Centers for Disease Control MMWR Office, v. 71, n. 21, p. 713–717, maio 2022. Disponível em: <https://doi.org/10.15585/mmwr.mm7121e1>.

CAVALCANTI, L. F.; BERTON, L. **Comparison of GANs for covid-19 x-ray classification**. In: *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021)*. [S.l.]: Sociedade Brasileira de Computação - SBC, 2021.

CHAKRABORTY, S.; MURALI, B.; MITRA, A. K. **An efficient deep learning model to detect COVID-19 using chest x-ray images**. *International Journal of Environmental Research and Public Health*, MDPI AG, v. 19, n. 4, p. 2013, fev. 2022. Disponível em: <https://doi.org/10.3390/ijerph19042013>.

CHLAP, P.; MIN, H.; VANDENBERG, N.; DOWLING, J.; HOLLOWAY, L.; HAWORTH, A. **A review of medical image data augmentation techniques for deep learning applications**. *J. Med. Imaging Radiat. Oncol.*, Wiley, v. 65, n. 5, p. 545–563, ago. 2021.

CHUNG, A.; WANG, L.; WONG, A.; LIN, Z. Q.; MCINNIS, P.; GUNRAJ, H. **Actualmed COVID-19 Chest X-ray Dataset Initiative**. [S.l.]: GitHub, 2020. <https://github.com/agchung/Actualmed-COVID-chestxray-datase>.

CHUNG, A.; WANG, L.; WONG, A.; LIN, Z. Q.; MCINNIS, P.; GUNRAJ, H. **Figure 1 COVID-19 Chest X-ray Dataset Initiative**. [S.l.]: GitHub, 2020. <https://github.com/agchung/Figure1-COVID-chestxray-dataset>.

CLAESSENS, Y.-E.; DEBRAY, M.-P.; TUBACH, F.; BRUN, A.-L.; RAMMAERT, B.; HAUSFATER, P.; NACCACHE, J.-M.; RAY, P.; CHOQUET, C.; CARETTE, M.-F.; MAYAUD, C.; LEPORT, C.; DUVAL, X. **Early chest computed tomography scan to assist diagnosis and guide treatment decision for suspected communityacquired pneumonia**. *American Journal of Respiratory and Critical Care Medicine*, American Thoracic Society, v. 192, n. 8, p. 974–982, out. 2015. Disponível em: <https://doi.org/10.1164/rccm.201501-0017oc>.

COHEN, J. P.; MORRISON, P.; DAO, L.; ROTH, K.; DUONG, T. Q.; GHASSEMI, M. **Covid-19 image data collection: Prospective predictions are the future**. *arXiv 2006.11988*, 2020. Disponível em: <https://github.com/ieee8023/covid-chestxray-dataset>.

DAHL, G. E.; YU, D.; DENG, L.; ACERO, A. **Large vocabulary continuous speech recognition with context-dependent DBN-HMMS**. IEEE, maio 2011. Disponível em: <https://doi.org/10.1109/icassp.2011.5947401>.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. **Imagenet: A large-scale hierarchical image database**. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255.

DENG, L. **The mnist database of handwritten digit images for machine learning research**. *IEEE Signal Processing Magazine*, IEEE, v. 29, n. 6, p. 141–142, 2012.

DIEBER, J.; KIRRANE, S. ***Why model why? Assessing the strengths and limitations of LIME***. 2020.

DILSIZIAN, S. E.; SIEGEL, E. L. **Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment**. *Current Cardiology Reports*, Springer Science and Business Media LLC, v. 16, n. 1, dez. 2013. Disponível em: <https://doi.org/10.1007/s11886-013-0441-8>.

DODGE, S.; KARAM, L. ***A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions***. 2017.

FANG, Y.; ZHANG, H.; XIE, J.; LIN, M.; YING, L.; PANG, P.; JI, W. S**ensitivity of chest CT for COVID-19: Comparison to RT-PCR**. *Radiology*, Radiological Society of North America (RSNA), v. 296, n. 2, p. E115–E117, ago. 2020. Disponível em: <https://doi.org/10.1148/radiol.2020200432>.

FAWCETT, T. **An introduction to roc analysis**. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016786550500303X>.

GOMES ALEX SANDRO; GOMES, C. R. A. **Classificação dos tipos de pesquisa em informática na educação**. In: JAQUES PATRíCIA AUGUSTIN; PIMENTEL, M. S. S. B. I. (Ed.). *Metodologia de Pesquisa Científica em Informática na Educação: Concepção de Pesquisa*. Porto Alegre: SBC, 2020, (Série Metodologia de Pesquisa em Informática na Educação, v. 1). cap. 4.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: GHAHRAMANI, Z.; WELLING, M.; CORTES, C.; LAWRENCE, N.; WEINBERGER, K. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2014. v. 27. Disponível em: <https://proceedings.neurips.cc/paper/2014/file/ 5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.

GUAN, W. jie; NI, Z. yi; HU, Y.; LIANG, W. hua; OU, C. quan; HE, J. xing; LIU, L.; SHAN, H.; LEI, C. liang; HUI, D. S.; DU, B.; LI, L. juan; ZENG, G.; YUEN, K.-Y.; CHEN, R. chong; TANG, C. li; WANG, T.; CHEN, P. yan; XIANG, J.; LI, S. yue; WANG, J. lin; LIANG, Z. jing; PENG, Y. xiang; WEI, L.; LIU, Y.; HU, Y. hua; PENG, P.; WANG, J. ming; LIU, J. yang; CHEN, Z.; LI, G.; ZHENG, Z. jian; QIU, S. qin; LUO, J.; YE, C. jiang; ZHU, S. yong; ZHONG, N. shan. **Clinical characteristics of coronavirus disease 2019 in china**. *New England Journal*

of Medicine, Massachusetts Medical Society, v. 382, n. 18, p. 1708–1720, abr. 2020. Disponível em: <https://doi.org/10.1056/nejmoa2002032>.

HAYKIN, S. O. **Neural Networks and Learning Machines**. 3. ed. Upper Saddle River, NJ: Pearson, 2008.

HINTON, G.; DENG, L.; YU, D.; DAHL, G.; MOHAMED, A. rahman; JAITLY, N.; SENIOR, A.; VANHOUCKE, V.; NGUYEN, P.; SAINATH, T.; KINGSBURY, B. **Deep neural networks for acoustic modeling in speech recognition**: The shared views of four research groups. *IEEE Signal Processing Magazine*, Institute of Electrical and Electronics Engineers (IEEE), v. 29, n. 6, p. 82–97, nov. 2012. Disponível em: <https://doi.org/10.1109/msp.2012.2205597>.

HODGES, A. ***Alan Turing: the enigma***. New York: Simon and Schuster, 1983. ISBN 978-0-671-49207-6 978-0-671-52809-6.

HUANG, C.; WANG, Y.; LI, X.; REN, L.; ZHAO, J.; HU, Y.; ZHANG, L.; FAN, G.; XU, J.; GU, X.; CHENG, Z.; YU, T.; XIA, J.; WEI, Y.; WU, W.; XIE, X.; YIN, W.; LI, H.; LIU, M.; XIAO, Y.; GAO, H.; GUO, L.; XIE, J.; WANG, G.; JIANG, R.; GAO, Z.; JIN, Q.; WANG, J.; CAO, B. **Clinical features of patients infected with 2019 novel coronavirus in wuhan, china**. *The Lancet*, Elsevier BV, v. 395, n. 10223, p. 497–506, fev. 2020. Disponível em: <https://doi.org/10.1016/s0140-6736(20)30183-5>.

HUANG, G.; LIU, Z.; MAATEN, L. van der; WEINBERGER, K. Q. **Densely connected convolutional networks**. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1608. 06993>.

JAEGER, S.; CANDEMIR, S.; ANTANI, S.; WÁNG, Y.-X. J.; LU, P.-X.; THOMA, G. **Two public chest x-ray datasets for computer-aided screening of pulmonary diseases**. *Quant. Imaging Med. Surg.*, v. 4, n. 6, p. 475–477, dez. 2014.

JSRT, J. S. o. R. T. **JSRT Database**. 1998. <http://db.jsrt.or.jp/eng.php>. Acesso em: 2021-12-20.

KATUWAL, G. J.; CHEN, R. **Machine Learning Model Interpretability for Precision Medicine**. 2016.

KERMANY, D. **Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. Mendeley**, 2018. Disponível em: <https://data.mendeley.com/datasets/rscbjbr9sj/2>.

KHAN, A. I.; SHAH, J. L.; BHAT, M. M. **CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images**. *Computer Methods and Programs in Biomedicine*, Elsevier BV, v. 196, p. 105581, nov. 2020. Disponível em: <https://doi.org/10.1016/j.cmpb.2020.105581>.

KHAN, S.; ALI, A.; SIDDIQUE, R.; NABI, G. **Novel coronavirus is putting the whole world on alert**. *Journal of Hospital Infection*, Elsevier BV, v. 104, n. 3, p. 252–253, mar 2020. Disponível em: <https://doi.org/10.1016%2Fj.jhin.2020.01.019>.

KIM, M.; YUN, J.; CHO, Y.; SHIN, K.; JANG, R.; BAE, H.-J.; KIM, N. **Deep learning in medical imaging**. *Neurospine*, The Korean Spinal Neurosurgery Society, v. 16, n. 4, p. 657–668, dez. 2019.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. **Imagenet classification with deep convolutional neural networks**. In: . [S.l.: s.n.], 2012. p. 1097–1105.

LEARNING, G. *Everything you need to know about VGG16*. Me-dium, 2021. Disponível em: <https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918>.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep learning**. *Nature*, Springer Science and Business Media LLC, v. 521, n. 7553, p. 436–444, maio 2015. Disponível em: <https://doi.org/10.1038/nature14539>.

LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. **Backpropagation applied to handwritten zip code recognition**. *Neural Computation*, MIT Press - Journals, v. 1, n. 4, p. 541–551, dez. 1989. Disponível em: <https://doi.org/10.1162/neco.1989.1.4.541>.

LI, Q.; GUAN, X.; WU, P.; WANG, X.; ZHOU, L.; TONG, Y.; REN, R.; LEUNG, K. S.; LAU, E. H.; WONG, J. Y.; XING, X.; XIANG, N.; WU, Y.; LI, C.; CHEN, Q.; LI, D.; LIU, T.; ZHAO, J.; LIU, M.; TU, W.; CHEN, C.; JIN, L.; YANG, R.; WANG, Q.; ZHOU, S.; WANG, R.; LIU, H.; LUO, Y.; LIU, Y.; SHAO, G.; LI, H.; TAO, Z.; YANG, Y.; DENG, Z.; LIU, B.; MA, Z.; ZHANG, Y.; SHI, G.; LAM, T. T.; WU, J. T.; GAO, G. F.; COWLING, B. J.; YANG, B.; LEUNG, G. M.; FENG, Z. **Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia**. *New England Journal of Medicine*, Massachusetts Medical Society, v. 382, n. 13, p. 1199–1207, mar. 2020. Disponível em: <https://doi.org/10.1056/nejmoa2001316>.

LI, Z.; DONG, M.; WEN, S.; HU, X.; ZHOU, P.; ZENG, Z. **Clu-cnns: Object detection for medical images**. *Neurocomputing*, v. 350, p. 53–59, 2019. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231219305521>.

LIU, H.; LIU, F.; LI, J.; ZHANG, T.; WANG, D.; LAN, W. **Clinical and CT imaging features of the COVID-19 pneumonia**: Focus on pregnant women and children. *Journal of Infection*, Elsevier BV, v. 80, n. 5, p. e7–e13, maio 2020. Disponível em: <https://doi.org/10.1016/j.jinf.2020.03.007>.

LOEY, M.; SMARANDACHE, F.; KHALIFA, N. E. M. **Within the lack of chest COVID-19 x-ray datase**t: A novel detection model based on GAN and deep transfer learning. *Symmetry*,

MDPI AG, v. 12, n. 4, p. 651, abr. 2020. Disponível em: <https://doi.org/10.3390/sym12040651>.

MAIOR, C. B. S.; SANTANA, J. M. M.; LINS, I. D.; MOURA, M. J. C. **Convolutional neural network model based on radiological images to support COVID-19 diagnosis**: Evaluating database biases. *PLOS ONE*, Public Library of Science (PLoS), v. 16, n. 3, p. e0247839, mar. 2021. Disponível em: <https://doi.org/10.1371/journal.pone.0247839>.

MATHIEU EDOUARD; RITCHIE, H. O.-O. E. R. M. H. J. A. C. G. C. R.-G. L. **A global database of covid-19 vaccinations**. *Nature Human Behaviour*, July 2021. Disponível em: <https://doi.org/10.1038/s41562-021-01122-8>.

MIKOLAJCZYK, A.; GROCHOWSKI, M. **Data augmentation for improving deep learning in image classification problem**. IEEE, maio 2018.

MINAEE, S.; BOYKOV, Y.; PORIKLI, F.; PLAZA, A.; KEHTARNAVAZ, N.; TERZOPOULOS, D. **Image segmentation using deep learning**: A survey. jan. 2020.

MURPHY, K. P. ***Probabilistic Machine Learning****: An introduction*. MIT Press, 2022. Disponível em: <probml.ai>.

NEFOUSSI, S.; AMAMRA, A.; AMAROUCHE, I. A. **A comparative study of deep learning networks for COVID-19 recognition in chest x-ray images**. In: *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*. IEEE, 2021. Disponível em: <https://doi.org/10.1109/ihsh51661. 2021.9378703>.

NERI, M.; OSORIO, M. C. *Centro de Políticas sociais*. FGV Social, 2022. Disponível em: <https://www.cps.fgv.br/cps/bd/docs/FGV_Social_Neri_ RetornoParaEscolaJornadaPandemia.pdf>.

NG, M.-Y.; LEE, E. Y. P.; YANG, J.; YANG, F.; LI, X.; WANG, H.; LUI, M. M. sze; LO, C. S.-Y.; LEUNG, B.; KHONG, P.-L.; HUI, C. K.-M.; YUEN, K. yung; KUO, M. D. **Imaging profile of the COVID-19 infection**: Radiologic findings and literature review. *Radiology: Cardiothoracic Imaging*, Radiological Society of North America (RSNA), v. 2, n. 1, p. e200034, fev. 2020. Disponível em: <https://doi.org/10.1148/ryct.2020200034>.

NIELSEN, M. **Neural Networks and Deep Learning**. Determination Press, 2015. Disponível em: <https://books.google.com.br/books?id=STDBswEACAAJ>.

NIKOLAOU, V.; MASSARO, S.; FAKHIMI, M.; STERGIOULAS, L.; GARN, W. **COVID-19 diagnosis from chest x-rays**: developing a simple, fast, and accurate neural network. *Health*

Information Science and Systems, Springer Science and Business Media LLC, v. 9, n. 1, out. 2021. Disponível em: <https://doi.org/10.1007/ s13755-021-00166-4>.

OMS. **Who coronavirus (COVID-19) dashboard**. World Health Organization, 2022. Disponível em: <https://covid19.who.int/>.

OSMAN, A. A.; DAAJANI, M. M. A.; ALSAHAFI, A. J. **Re-positive coronavirus disease 2019 PCR test**: could it be a reinfection? *New Microbes New Infect.*, Elsevier BV, v. 37, n. 100748, p. 100748, set. 2020.

PEREZ, L.; WANG, J. **The effectiveness of data augmentation in image classification using deep learning**. arXiv, 2017.

POWERS, D. M. W. **Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation**. 2020.

PUNN, N. S.; AGARWAL, S. **Automated diagnosis of COVID-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks**. *Applied Intelligence*, Springer Science and Business Media LLC, v. 51, n. 5, p. 2689–2702, out. 2020. Disponível em: <https://doi.org/10.1007/s10489-020-01900-3>.

RADFORD, A.; METZ, L.; CHINTALA, S. **Unsupervised representation learning with deep convolutional generative adversarial networks**. nov. 2015.

RAHMAN, T.; CHOWDHURY, M.; KHANDAKAR, A. **Actualmed COVID-19 Chest X-ray Dataset Initiative**. [S.l.]: Kaggle, 2021. <https://www.kaggle.com/datasets/ tawsifurrahman/covid19-radiography-database>.

RAJKUMAR, S.; RAJARAMAN, P. V.; MEGANATHAN, H. S.; SAPTHAGIRIVASAN, V.; TEJASWINEE, K.; ASHWIN, R. COVID-DETECT: A DEEP LEARNING APPROACH FOR CLASSIFICATION OF COVID-19 PNEUMONIA FROM LUNG SEGMENTED CHEST x-RAYS. **Biomedical Engineering: Applications, Basis and Communications**, National Taiwan University, v. 33, n. 02, p. 2150010, dez. 2020. Disponível em: <https://doi.org/10.4015/s1016237221500101>.

RAJPURKAR, P.; IRVIN, J.; ZHU, K.; YANG, B.; MEHTA, H.; DUAN, T.; DING, D.; BAGUL, A.; LANGLOTZ, C.; SHPANSKAYA, K. et al. **Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning**. *arXiv preprint arXiv:1711.05225*, 2017.

RAJPURKAR, P.; IRVIN, J.; ZHU, K.; YANG, B.; MEHTA, H.; DUAN, T.; DING, D.; BAGUL, A.; LANGLOTZ, C.; SHPANSKAYA, K.; LUNGREN, M. P.; NG, A. Y. **CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning**. arXiv, 2017.

REYNOLDS, A. H. **Convolutional Neural Networks (cnns)**. 2017. Disponível em: <https://anhreynolds.com/blogs/cnn.html>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. **"why should I trust you?"**: Explaining the predictions of any classifier. arXiv, 2016.

ROCHA, L. **Falta de testes para covid: Dependência de Importação de insumos ajuda a explicar**. 2022. Disponível em: <https://www.cnnbrasil.com.br/saude/falta-de-testes-para-covid-dependencia-de-importacao-de-insumos-ajuda-a-explicar/>.

RODRIGUES, R. S.; RIBEIRO, G. M.; BARRETO, M. M.; ZIN, W. A.; TOLEDO-MENDES, J. de; MARTINS, P. A. G.; ALMEIDA, S. A. de; BASÍLIO, R.; MARTINS-GONÇALVES, R.; HOTTZ, E. D.; BOZZA, P. T.; BOZZA, F. A.; CARVALHO, A. R. S.; CASTRO, P. H. R. de. **Increased lung immune metabolic activity in COVID-19 survivors**. Clinical Nuclear Medicine, Ovid Technologies (Wolters Kluwer Health), Publish Ahead of Print, ago. 2022. Disponível em: <https://doi.org/10.1097/rlu.0000000000004376>.

RONNEBERGER, O.; FISCHER, P.; BROX, T. **U-Net: Convolutional networks for biomedical image segmentation**. Maio 2015.

RSNA, R. S. o. N. A. **RSNA Pneumonia Detection Challenge**. [S.l.]: Kaggle, 2018. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data?select=stage_ 2_detailed_class_info.csv>, note = Accessed: 2021-12-10.

RSNA, R. S. o. N. A. **RSNA International COVID19 Open Radiology Database (RICORD)**. 2020.
://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70230281.

RUBIN, G. D.; RYERSON, C. J.; HARAMATI, L. B.; SVERZELLATI, N.; KANNE, J. P.; RAOOF, S.; SCHLUGER, N. W.; VOLPI, A.; YIM, J.-J.; MARTIN, I. B.; ANDERSON, D. J.; KONG, C.; ALTES, T.; BUSH, A.; DESAI, S. R.; GOLDIN, J.; GOO, J. M.; HUMBERT, M.; INOUE, Y.; KAUCZOR, H.-U.; LUO, F.; MAZZONE, P. J.; PROKOP, M.; REMY-JARDIN, M.; RICHELDI, L.; SCHAEFER-PROKOP, C. M.; TOMIYAMA, N.; WELLS, A. U.; LEUNG, A. N. **The role of chest imaging in patient management during the COVID-19 pandemic**. *Chest*, Elsevier BV, v. 158, n. 1, p. 106–116, jul. 2020. Disponível em: <https://doi.org/10.1016/j.chest.2020.04.003>.

SAUDE, M. **Painel coronavírus**. 2022. Disponível em: <https://covid.saude.gov.br/>.

SHEN, L.; MARGOLIES, L. R.; ROTHSTEIN, J. H.; FLUDER, E.; MCBRIDE, R.; SIEH, W. **Deep learning to improve breast cancer detection on screening mammography**. *Scientific Reports*, Springer Science and Business Media LLC, v. 9, n. 1, ago. 2019. Disponível em: <https://doi.org/10.1038/s41598-019-48995-4>.

SHORTEN, C.; KHOSHGOFTAAR, T. M. **A survey on image data augmentation for deep learning**. *J. Big Data*, Springer Science and Business Media LLC, v. 6, n. 1, dez. 2019.

SIDDIQUE, N.; PAHEDING, S.; ELKIN, C. P.; DEVABHAKTUNI, V. **U-net and its variants for medical image segmentation**: A review of theory and applications. *IEEE Access*, v. 9, p. 82031–82057, 2021.

SIMONYAN, K.; ZISSERMAN, A. **Very deep convolutional networks for large-scale image recognition**. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1409.1556>.

SULTANA, F.; SUFIAN, A.; DUTTA, P. **Evolution of image segmentation using deep convolutional neural network**: A survey. jan. 2020.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. **Inception-v4, inceptionresnet and the impact of residual connections on learning**. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1602.07261>.

TATTI, N. **Maintaining auc and h-measure over time**. *Machine Learning*, 12 2021.

TEIXEIRA, L. O.; PEREIRA, R. M.; BERTOLINI, D.; OLIVEIRA, L. S.; NANNI, L.; CAVALCANTI, G. D. C.; COSTA, Y. M. G. **Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest x-ray images**. *Sensors*, MDPI AG, v. 21, n. 21, p. 7116, out. 2021. Disponível em: <https://doi.org/10.3390/s21217116>.

UCAR, F.; KORKMAZ, D. **COVIDiagnosis-net: Deep bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from x-ray images**. *Medical Hypotheses*, Elsevier BV, v. 140, p. 109761, jul. 2020. Disponível em: <https: //doi.org/10.1016/j.mehy.2020.109761>.

UMER, M.; ASHRAF, I.; ULLAH, S.; MEHMOOD, A.; CHOI, G. S. **COVINet: a convolutional neural network approach for predicting COVID-19 from chest x-ray images**. *Journal of Ambient Intelligence and Humanized Computing*, Springer Science and Business Media LLC, v. 13, n. 1, p. 535–547, jan. 2021. Disponível em: <https://doi.org/10.1007/s12652-021-02917-3>.

VAYÁ, M. d. l. I.; SABORIT, J. M.; MONTELL, J. A.; PERTUSA, A.; BUSTOS, A.; CAZORLA, M.; GALANT, J.; BARBER, X.; OROZCO-BELTRÁN, D.; GARCÍAGARCÍA, F.; CAPARRÓS, M.; GONZÁLEZ, G.; SALINAS, J. M. **BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients**. arXiv, 2020.

WANG, L.; LIN, Z. Q.; WONG, A. **COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images**. *Scientific Reports*, Springer Science and Business Media LLC, v. 10, n. 1, nov. 2020. Disponível em: <https://doi.org/10.1038/s41598-020-76550-z>.

WANG, R.; LEI, T.; CUI, R.; ZHANG, B.; MENG, H.; NANDI, A. K. **Medical image segmentation using deep learning**: A survey. set. 2020.

WASON, R. **Deep learning: Evolution and expansion**. *Cogn. Syst. Res.*, Elsevier BV, v. 52, p. 701–708, dez. 2018.

WAZLAWICK, R. S. **Metodologia de pesquisa para ciência da computação.** Rio de Janeiro: Elsevier, 2014.

WHO. **WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 — who.int**. 2020. <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>. [Accessed 02-Aug-2022].

WONG, H. Y. F.; LAM, H. Y. S.; FONG, A. H.-T.; LEUNG, S. T.; CHIN, T. W.-Y.; LO, C. S. Y.; LUI, M. M.-S.; LEE, J. C. Y.; CHIU, K. W.-H.; CHUNG, T. W.-H.; LEE, E. Y. P.; WAN, E. Y. F.; HUNG, I. F. N.; LAM, T. P. W.; KUO, M. D.; NG, M.-Y. **Frequency and distribution of chest radiographic findings in patients positive for COVID-19**. *Radiology*, Radiological Society of North America (RSNA), v. 296, n. 2, p. E72–E78, ago. 2020. Disponível em: <https://doi.org/10.1148/radiol.2020201160>.

XIE, X.; ZHONG, Z.; ZHAO, W.; ZHENG, C.; WANG, F.; LIU, J. **Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: Relationship to negative RT-PCR testing**. *Radiology*, Radiological Society of North America (RSNA), v. 296, n. 2, p. E41–E45, ago. 2020. Disponível em: <https://doi.org/10.1148/radiol.2020200343>.

YU, F.; ZHANG, Y.; SONG, S.; SEFF, A.; XIAO, J. **Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop**. *CoRR*, abs/1506.03365, 2015. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr1506.html#YuZSSX15>.